Ludwig-Maximilians-Universität

Department of Statistics

Master's Thesis

Data dredging in ranking analyses with focus on biomedical applications

Author: Christina Nießl

Supervisor: Prof. Dr. Anne-Laure Boulesteix

Submission date: 27. November 2019

Abstract

Ranking analyses are an essential step in biomedical studies investigating high-dimensional molecular data. In general, there are many ranking methods available among which researchers can choose. Moreover, ranking analyses involving high-dimensional data are generally highly unstable in the sense that a small modification of the ranking method can lead to a completely different ordering of the variables. This increases the temptation to apply several ranking methods until one of them returns a satisfying result. The strategy of choosing the analysis approach based on its results is commonly referred to as data dredging and is known to lead to substantial optimistic biases. To raise awareness of this practice, it could be useful to give researchers a concrete idea of how unstable ranking results are with respect to data dredging. This certain type of variability can be referred to as data dredging potential. In this thesis, a framework for the quantification of data dredging potential in ranking analyses is provided. The proposed framework is illustrated in the context of genes rankings using simulated and real data sets. The findings suggest that many ranking results show a data dredging potential that can be considered as problematic.

Contents

List of Figures

List of Tables

1. Introduction

In many research fields, it is necessary to rank variables according to their importance for an outcome of interest. In biomedical sciences, an application field of rankings is the identification of molecular biomarkers (Klau et al., 2019). In simple terms, molecular biomarkers are molecules such as genes, proteins or metabolites that can be used for disease diagnosis and prognosis, prediction of therapeutic responses or therapeutic development (Hu et al., 2011). As a consequence of the advance of high-throughput technologies in the last two decades, molecular data sets are usually high-dimensional in the sense that they contain thousands of measurements for each observation (Boulesteix et al., 2017).

In general, variable rankings are performed using a ranking criterion that evaluates the degree of correlation between each variable and the outcome of interest (Dessì et al., 2013). Like in other statistical analyses, generating variable rankings requires several decisions regarding the analysis approach. This does not only include the ranking criterion itself but also, for instance, data preparation steps or the choice of parameter values (Klau et al., 2019). All choices regarding a statistical analysis approach are generally referred to as *researcher degrees of freedom* (Simmons et al., 2011).

While the multitude of analysis approaches is an issue that concerns all research fields, it is particularly relevant for variable rankings and other statistical analyses involving molecular data. This has several reasons: Firstly, it has been recognized that many traditional analysis approaches are inappropriate considering the complexity and high-dimensionality of molecular data. Therefore, alternative approaches have been developed, which has led to an increase in researcher degrees of freedom. At the same time, there is a lack of guidelines supporting the choice of statistical analysis approaches. Moreover, the high-dimensionality of molecular data leads to a high variability in results. Regarding ranking analyses, this means that minor changes in the sample or small modifications of the ranking criterion can result in a completely different ordering of the variables. (Boulesteix and Slawski, 2009; Boulesteix et al., 2017)

As a consequence, ranking analyses involving high-dimensional molecular data can be expected to leave much room for *data dredging*, which is the (conscious or subconscious) strategy of applying several analysis strategies and only reporting the "best-looking" result (Ioannidis, 2005; Boulesteix et al., 2017). In the context of ranking analyses, the "best-looking" result could be for instance a top-rank for the researcher's "favourite" variable (e.g. a variable that is expected to be relevant due to biological knowledge or because previous studies have shown its relevance) (Boulesteix and Slawski, 2009). Even if all

methods are conceivable from a theoretical and practical point of view, data dredging can lead to a substantial optimistic bias and false research findings (Ioannidis, 2005).
Although data dredging has gained more attention in recent years (e.g. Szucs, 2016; Marín-Franch, 2018), it might still be an abstract concept for some researchers. Thus, for a researcher that has performed ranking analyses using several analysis approaches, it might be helpful to know how much room the ranking results actually leave for data dredging. The variability in ranking results with respect to data dredging will be referred to as *data dredging potential*. In the bioinformatics literature, several measurements have been proposed to assess the stability of ranking lists (Boulesteix and Slawski, 2009). However, they do not allow to quantify the variability in ranking results with respect to data dredging. Thus, the aim of this thesis is to provide a framework that allows to quantify the data dredging potential in ranking results in a simple and comprehensive way.

The thesis is structured as follows: Chapter 2 introduces a framework for the quantification of data dredging potential in ranking results. The proposed framework is illustrated in the context of gene rankings using different data sets: In Chapter 3, the framework is applied to simulated data with different parameter settings and a binary outcome. In Chapter 4 and 5, the proposed framework is illustrated using real data with binary and survival outcome. Chapter 6 summarizes the main findings, discusses the strengths and limitations of the framework and gives an outlook.

The results shown in this thesis were generated using version 3.5.3 of the software `R` (R Core Team, 2019). In addition to the packages cited in the text, the `R`-packages `ggplot2` (Wickham, 2016), `reshape2` (Wickham, 2007), `latex2exp` (Meschiari, 2015), `mvtnorm` (Genz et al., 2019), `dplyr` (Wickham et al., 2019), `stringr` (Wickham, 2019) and `xtable` (Dahl et al., 2019) were used.

2. Framework

This chapter provides a framework for the quantification of data dredging potential in ranking results. To begin with, Section 2.1 formalises data dredging in the context of ranking analyses. Section 2.2 then introduces a measure that allows to quantify the data dredging potential of each ranking result. Finally, two possible applications in practice are considered in Section 2.3.

2.1. Formalisation of data dredging in the context of rankings

Data dredging is the conscious or subconscious strategy of applying several analysis methods and only reporting the method that yields the "best-looking" result (Ioannidis, 2005; Boulesteix et al., 2017). In this section, data dredging will be formalised in the context of ranking analyses.

Let $X_1, \ldots, X_p$ be the variables that a researcher wants to rank according to their association with an outcome Y. For example, the researcher might be interested in knowing which genes $X_1, \ldots, X_p$ can be used to diagnose a certain disease. In this case, Y is a binary variable (disease is present/not present). However, it could also be, for instance, a categorial variable or a censored survival time. Moreover, let $\mathcal{D}$ denote a data set that contains n observations of the variables $X_1, \ldots, X_p$ and Y. (Boulesteix and Slawski, 2009)

In this work, a *ranking* of the variables $X_1, \ldots, X_p$ is defined as an ordered list where the p variables appear in descending order of relevance for the outcome. In ranking analyses, the relevance of a variable is determined by a *ranking criterion*, which measures the degree of correlation between each variable and the outcome in the given data set $\mathcal{D}$. Based on the value of the ranking criterion, a *rank* $r_j \in \{1, \ldots, p\}$ is assigned to each variable X_j $(j = 1, \ldots, p)$. A small value of r_j indicates a strong association between X_j and Y, either positive or negative. For example, $r_j = 1$ means that variable X_j is identified as the most relevant variable. (Dessì et al., 2013; Boulesteix and Slawski, 2009)

Depending on the type of outcome, there are usually many ranking methods available which consequently lead to different variable rankings. In the case of data dredging, the researcher does not choose one of them but tries out several ranking methods. Thus, let m be the number of ranking methods that were applied to a given data set $\mathcal{D}$. Note that the term "method" does not only refer to the ranking criterion itself but to all choices concerning the analysis approach, e.g. potential tuning parameters (Klau et al., 2019). Furthermore, let $\boldsymbol{r}_j = (r_{j1} \quad r_{j2} \quad \cdots \quad r_{jm})$ be a vector that contains the resulting m ranks of variable X_j, where r_{jk} is the rank assigned to X_j by method k $(k = 1, \ldots, m)$. The

vector $\boldsymbol{r}_j$ will be referred to as *ranking result* of variable X_j.
To further formalise data dredging, it is required to define what is considered as the "best-looking" result in the context of rankings. In this work, it is assumed that the researcher has a "favourite" variable that he/she expects to be identified as relevant (e.g. because a previous study has shown its relevance or because of biological knowledge) (Boulesteix and Slawski, 2009). Then, the best-looking result is the smallest rank that was assigned to the favourite variable by one of the m applied methods. Consequently, the method that yields the best-looking result varies depending on which variable is the researcher's favourite variable. Formally speaking, define the *best rank* of variable X_j as

$$r_j^{best} := r_{j\,\min\{k:\, r_k = \min \boldsymbol{r}_j\}}. \tag{2.1}$$

Based on this definition, $\boldsymbol{r}_j^{-best}$ denotes the vector of length $m-1$ that contains all ranks of variable X_j except r_j^{best}. Defining r_j^{best} as in Equation (2.1) (and not just as the minimum of $\boldsymbol{r}_j$) ensures that $\boldsymbol{r}_j^{-best}$ contains $m-1$ ranks even if more than one method yields the smallest rank. It would however also be possible to define r_j^{best} as $r_{j\,\max\{k:\, r_k = \min \boldsymbol{r}_j\}}$. As an example, let $\boldsymbol{r}_j = (3 \quad 5 \quad 3 \quad 6)$ be the ranking result of variable X_j. In this case, the best rank is $r_j^{best} = r_{j1} = 3$ and the vector containing all ranks except r_j^{best} is $\boldsymbol{r}_j^{-best} = (r_{j2} \quad r_{j3} \quad r_{j4}) = (5 \quad 3 \quad 6)$.

Let X_{j*} be the researcher's favourite variable and $\boldsymbol{r}_{j*}$ the ranks that were assigned to X_{j*} by applying m methods on the same data set $\mathcal{D}$. There are two questions that may arise in this context: (i) how much room does ranking result $\boldsymbol{r}_{j*}$ leave for data dredging, and (ii) what are the actual consequences if the researcher decides to report only the best-looking result (i.e. r_{j*}^{best}) and ignores the other results (i.e. $\boldsymbol{r}_{j*}^{-best}$)? Based on the definitions given above, it is now possible to address these questions from a statistical perspective.

First, the question concerning the consequences of data dredging is discussed. It is commonly known that data dredging can lead to a substantial optimistic bias (Ioannidis, 2005), which will be referred to as *data dredging bias*. For instance, regarding high-dimensional class prediction, Boulesteix and Strobl (2009) have quantitatively assessed the bias in prediction error estimation that is caused by selecting the classifier a posteriori.
In the context of ranking analyses, an intuitive way to quantify the data dredging bias would be to compare r_{j*}^{best} with the true rank of variable X_{j*}. However, unless the data are simulated, the true rank is typically not known (Boulesteix and Slawski, 2009). In practice, it is therefore not possible to quantitatively assess the data dredging bias of $\boldsymbol{r}_{j*}$.

Therefore, this thesis will address the question of how much room ranking results leave for data dredging. This certain type of variability in results will be referred to as *data dredging potential* and the following sections will provide an approach for its quantitative

assessment. In contrast to the data dredging bias, this does not require any knowledge about the true rank and is therefore more useful in practice.
The next section introduces a measure that quantifies the data dredging potential of each ranking result $\boldsymbol{r}_j$. Section 2.3 will then show possible applications in practice. This includes the comparison of the data dredging potential of $\boldsymbol{r}_{j*}$ with the other ranking results and the assessment of the overall data dredging potential. The latter might be useful even if there is no specific favourite variable since it gives the researcher an overview about how (un)stable the ranking results are with respect to data dredging.

2.2. Quantification of data dredging potential

This section provides an approach to quantify the data dredging potential of a single ranking result $\boldsymbol{r}_j$. Section 2.2.1 examines the attributes of $\boldsymbol{r}_j$ that determine its data dredging potential and deduces several aspects that a measure for data dredging potential should satisfy. Based on these considerations, Section 2.2.2 introduces a suitable measure, which is then modified in Section 2.2.3 to meet all requirements.

2.2.1. Important aspects

Essentially, there are two attributes of $\boldsymbol{r}_j$ that determine its data dredging potential. To illustrate this, let $\boldsymbol{r}_j = (3 \quad 9 \quad 10 \quad 15)$ and $\boldsymbol{r}_{j'} = (3 \quad 9 \quad 10 \quad 12)$ be the ranking results of the variables X_j and $X_{j'}$. In both cases, the best rank is equal to 3 ($r_j^{best} = r_{j'}^{best} = 3$). However, because $r_{j4} > r_{j'4}$, only reporting the best rank intuitively seems to be more problematic for $\boldsymbol{r}_j = (3 \quad 9 \quad 10 \quad 15)$ than for $\boldsymbol{r}_{j'} = (3 \quad 9 \quad 10 \quad 12)$. In other words, $\boldsymbol{r}_j$ has a higher data dredging potential than $\boldsymbol{r}_{j'}$. Thus, to assess the data dredging potential of $\boldsymbol{r}_j$, the range and distribution of $\boldsymbol{r}_j^{-best}$ with respect to r_j^{best} should be considered. However, this is not the only attribute of $\boldsymbol{r}_j$ that determines its data dredging potential. Consider for example the ranking results $\boldsymbol{r}_j = (1 \quad 7 \quad 8 \quad 10)$ and $\boldsymbol{r}_{j'} = (101 \quad 107 \quad 108 \quad 110)$. Although the differences between the ranks in $\boldsymbol{r}_j$ and $\boldsymbol{r}_{j'}$ are equal (namely $6 - 1 - 2$), $\boldsymbol{r}_j$ has a higher data dredging potential than $\boldsymbol{r}_{j'}$. This is because $r_j^{best} = 1$ identifies X_j as the most important variable for Y, which will be more relevant for the researcher than $r_{j'}^{best} = 101$. Therefore, the second important attribute of $\boldsymbol{r}_j$ that should be taken into account to assess the data dredging potential is the actual value of r_j^{best}.
Based on these considerations, a ranking result with the highest possible data dredging potential is a ranking result where r_j^{best} is equal to 1 and the other $m - 1$ ranks in $\boldsymbol{r}_j^{-best}$ are equal to the highest possible rank p. For instance, the ranking results with the highest possible data dredging potential for $p = 1000$ and $m = 3$ are $\boldsymbol{r}_j = (1 \quad 1000 \quad 1000)$, $\boldsymbol{r}_j = (1000 \quad 1 \quad 1000)$ and $\boldsymbol{r}_j = (1000 \quad 1000 \quad 1)$. Because such a result is the most problematic ranking result concerning data dredging, it will be referred to as *worst case result*. Conversely, a *best case result* would be a ranking result that has no data dredging poten-

tial. This is obviously the case if all methods yield the same rank, e.g. $\boldsymbol{r}_j = (5 \quad 5 \quad 5 \quad 5)$ or $\boldsymbol{r}_j = (1 \quad 1 \quad 1 \quad 1)$.

To sum up, a measure that quantifies the data dredging potential of $\boldsymbol{r}_j$ should take the distribution of $\boldsymbol{r}_j^{-best}$ with respect to r_j^{best} and the actual value of r_j^{best} into account. Moreover, its minimum and maximum values should ideally correspond to the worst and best case ranking results. In principle, such a measure would be a stability measure that quantifies the variability in $\boldsymbol{r}_j$ with respect to r_j^{best}.
Several stability measures for ranking lists can already be found in existing literature. For example, many approaches assess the similarity of ranking lists by considering their overlap. Other approaches use adaptions of distance measures like Spearman's correlation coefficient (Boulesteix and Slawski, 2009). For more information on stability measures for ranking lists see for example Boulesteix and Slawski (2009) and the references therein. However, to our knowledge, no stability measure exists that is appropriate for the quantification of the data dredging potential in $\boldsymbol{r}_j$. One reason for this is that many measures do not allow to compare rankings of $m > 2$ methods without considering them pairwise or comparing all rankings to a reference ranking list (Boulesteix and Slawski, 2009). Furthermore, they can not be applied to a single ranking result $\boldsymbol{r}_j$ and do not measure the stability with respect to r_j^{best}.
The next two sections will thus introduce a measure that takes all important aspects into account.

2.2.2. Distribution and range with respect to best rank

A possible way to quantify the data dredging potential of $\boldsymbol{r}_j$ is to define $h(\boldsymbol{r}_j)$ as the difference between the mean over all ranks in $\boldsymbol{r}_j$ and r_j^{best}:

$$h(\boldsymbol{r}_j) := \big(\frac{1}{m}\sum_{k=1}^{m} r_{jk}\big) - r_j^{best} = \bar{\boldsymbol{r}}_j - r_j^{best}. \tag{2.2}$$

The higher the ranks in $\boldsymbol{r}_j^{-best}$ with respect to r_j^{best}, the higher the value of $h(\boldsymbol{r}_j)$. Consequently, a high value of $h(\boldsymbol{r}_j)$ indicates a high data dredging potential of $\boldsymbol{r}_j$.
As an example, Table 2.1 shows three different ranking results and the corresponding values of $h(\boldsymbol{r}_j)$. Ranking result $\boldsymbol{r}_j = (3 \quad 9 \quad 10 \quad 15)$ has the highest ranking potential and accordingly yields the highest value of $h(\boldsymbol{r}_j)$.
Concerning the range of values, $h(\boldsymbol{r}_j)$ takes its minimum value if and only if r_j^{best} is equal to $\bar{\boldsymbol{r}}_j$, i.e. if all m methods yield the same rank. Conversely, $h(\boldsymbol{r}_j)$ takes its maximum value if and only if r_j^{best} is equal to 1 and all ranks in $\boldsymbol{r}_j^{-best}$ are equal to p. In this case,

Table 2.1.: Data dredging potential of three example ranking results measured by $h(\boldsymbol{r}_j)$.

$\boldsymbol{r}_j$	$\bar{\boldsymbol{r}}_j$	r_j^{best}	$h(\boldsymbol{r}_j)$
(3 5 10 12)	7.5	3	4.5
(3 9 10 12)	8.5	3	5.5
(3 9 10 15)	9.25	3	6.25

$h(\boldsymbol{r}_j)$ is equal to $\frac{1+(m-1)p}{m} - 1 = \frac{m-1}{m}(p-1)$. This means that

$$h(\boldsymbol{r}_j) \in [0, \frac{m-1}{m}(p-1)], \tag{2.3}$$

where the minimum and maximum values correspond to the best and worst case result of $\boldsymbol{r}_j$, respectively. Note that the maximum value of $h(\boldsymbol{r}_j)$ consequently depends on the number of methods and variables. However, this is reasonable because for instance, the worst case result for $m = 4$ and $p = 1000$ (e.g. $\boldsymbol{r}_j = (1 \quad 1000 \quad 1000 \quad 1000)$) seems to be more problematic than the worst case result for $m = 2$ and $p = 100$ (e.g. $\boldsymbol{r}_j = (1 \quad 100)$). This is reflected by the values of $h(\boldsymbol{r}_j)$, namely $h(1, 100) = 49.5$ and $h(1, 1000, 1000, 1000) = 749.25$. Thus, if $h(\boldsymbol{r}_j)$ would be normed to have the same range of values for every combination of p and m, comparison of $h(\boldsymbol{r}_j)$ for different m and p settings would get complicated.

Graphical interpretation

A useful feature of $h(\boldsymbol{r}_j)$ is its graphical interpretation. Let $f(s; \boldsymbol{r}_j)$ be a step function returning the proportion of ranks in $\boldsymbol{r}_j$ that are larger than s if $s \geqslant \min \boldsymbol{r}_j$:

$$f(s; \boldsymbol{r}_j) = \begin{cases} 0, & \text{if } s < \min \boldsymbol{r}_j, \\ \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}_{\{r_{jk} > s\}}, & \text{if } s \geqslant \min \boldsymbol{r}_j, \end{cases} \tag{2.4}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The area under $f(s; \boldsymbol{r}_j)$ equals $h(\boldsymbol{r}_j)$, which means that the higher the area under $f(s; \boldsymbol{r}_j)$, the higher the data dredging potential of ranking result $\boldsymbol{r}_j$.

Figure 2.1 displays $f(s; \boldsymbol{r}_j)$ for the same three example ranking results that were shown in Table 2.1. By comparing the areas under $f(s; \boldsymbol{r}_j)$, it can be seen that $\boldsymbol{r}_j = (3 \quad 9 \quad 10 \quad 15)$ has the highest data dredging potential.

The equality between the area under $f(s; \boldsymbol{r}_j)$ and $h(\boldsymbol{r}_j)$ can be shown by integrating

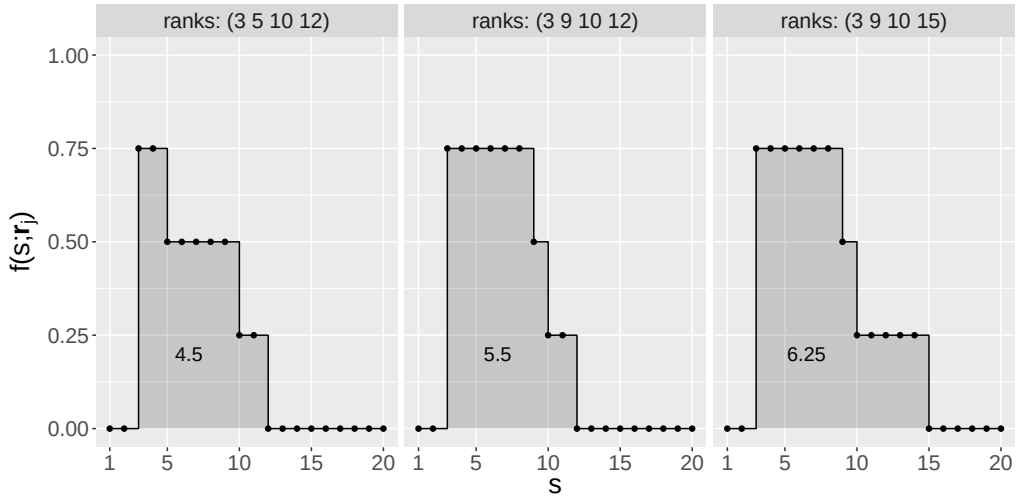


Figure 2.1.: Graphical interpretation of $h(\boldsymbol{r}_j)$ for three example ranking results.

$f(s;\boldsymbol{r}_j)$:

$$\begin{aligned}\int_{-\infty}^{\infty} f(s;\boldsymbol{r}_j)ds &= \int_{\min \boldsymbol{r}_j}^{\max \boldsymbol{r}_j} f(s;\boldsymbol{r}_j)ds \\ &= \int_{\min \boldsymbol{r}_j}^{\max \boldsymbol{r}_j} \frac{\sum_{k=1}^{m} \mathbb{1}_{\{r_{jk}>s\}}}{m} ds \\ &= \sum_{s=\min \boldsymbol{r}_j}^{\max \boldsymbol{r}_j} \frac{\#\{k : r_{jk} > s\}}{m}. \end{aligned} \tag{2.5}$$

Equation (2.5) shows that the area under $f(s;\boldsymbol{r}_j)$ can be calculated by adding the proportions of ranks in $\boldsymbol{r}_j$ larger than s for all $s = \min \boldsymbol{r}_j, \ldots, \max \boldsymbol{r}_j$. Assume now without loss of generality that the ranks in $\boldsymbol{r}_j$ are ordered such that $r_{j1} \leqslant r_{j2} \leqslant \cdots \leqslant r_{jm}$. In this case, Equation (2.5) can be written as

$$\begin{aligned}\sum_{s=r_j^{best}}^{r_j^{worst}} \frac{\#\{k : r_{jk} > s\}}{m} &= \frac{\sum_{k=1}^{m-1}(r_{j(k+1)} - r_{jk})(m-k)}{m} \\ &= \frac{\sum_{k=1}^{m-1} r_{j(k+1)}(m-k) - \sum_{k=1}^{m-1} r_{jk}(m-k)}{m} \\ &= \frac{\sum_{k=2}^{m} r_{jk}}{m} - \frac{r_{j1}(m-1)}{m} \\ &= \frac{\sum_{k=2}^{m} r_{jk}}{m} + \frac{r_{j1}}{m} - r_{j1} \\ &= \frac{\sum_{k=1}^{m} r_{jk}}{m} - r_{j1} \\ &= h(\boldsymbol{r}_j). \end{aligned} \tag{2.6}$$

This shows that $h(\boldsymbol{r}_j)$ is equal to the area under $f(s; \boldsymbol{r}_j)$.

Comparison with an alternative approach

Another possible approach for quantifying the data dredging potential is based on Ward's method, which is a hierarchical clustering procedure (Ward, 1963). In each step of agglomerative hierarchical clustering, Ward's method joins the two clusters A and B that minimise the increase in the sum of squared errors (SSE):

$$\begin{aligned} D(A,B) &= \sum_{i \in A \cup B} \|\boldsymbol{z}_i - \bar{\boldsymbol{z}}_{A\cup B}\|^2 - \Big(\sum_{i \in A} \|\boldsymbol{z}_i - \bar{\boldsymbol{z}}_A\|^2 + \sum_{i \in B} \|\boldsymbol{z}_i - \bar{\boldsymbol{z}}_B\|^2 \Big) \\ &= SSE_{A\cup B} - (SSE_A + SSE_B), \end{aligned} \tag{2.7}$$

where $\|\cdot\|$ is the Euclidean norm, $\boldsymbol{z}_i$ is an observation of the random vector $(Z_1, \ldots, Z_p)^\top$ and $\bar{\boldsymbol{z}}_A$, $\bar{\boldsymbol{z}}_B$ and $\bar{\boldsymbol{z}}_{A\cup B}$ are the centroids of clusters A, B and $A \cup B$, respectively (Rencher, 2002). It can be shown that the increase in SSE is equivalent to

$$D(A,B) = \frac{n_A n_B}{n_A + n_B} \|\bar{\boldsymbol{z}}_A - \bar{\boldsymbol{z}}_B\|^2 \tag{2.8}$$

where n_A, n_B and are the number of observations in cluster A and B, respectively (Rencher, 2002).

The concept of comparing SSEs cannot only be used in the clustering context to find an optimal grouping for observations but also to quantify the data dredging potential in ranking results. For this purpose, consider the ranks in $\boldsymbol{r}_j^{-best}$ as cluster A ($n_A = m - 1$) and r_j^{best} as cluster B ($n_B = 1$). An alternative measure for quantifying the data dredging potential of $\boldsymbol{r}_j$ can be defined as the increase in SSE by "adding" r_j^{best} to $\boldsymbol{r}_j^{-best}$:

$$h_2(\boldsymbol{r}_j) := D(\boldsymbol{r}_j^{-best}, r_j^{best}) = \frac{m-1}{m} (\overline{\boldsymbol{r}_j^{-best}} - r_j^{best})^2, \tag{2.9}$$

where $\overline{\boldsymbol{r}_j^{-best}}$ is the mean over all ranks in $\boldsymbol{r}_j^{-best}$. The higher the ranks in $\boldsymbol{r}_j^{-best}$ compared to r_j^{best}, the higher the increase in SSE. Consequently, a high value of $h_2(\boldsymbol{r}_j)$ indicates a high data dredging potential of $\boldsymbol{r}_j$.

Table 2.2 contains the values of $h_2(\boldsymbol{r}_j)$ for the three example ranking results that were shown in Table 2.1.

Table 2.2.: Data dredging potential of three example ranking results measured by $h_2(\boldsymbol{r}_j)$.

$\boldsymbol{r}_j$	$h_2(\boldsymbol{r}_j)$
(3 5 10 12)	27.00000
(3 9 10 12)	40.33333
(3 9 10 15)	52.08333

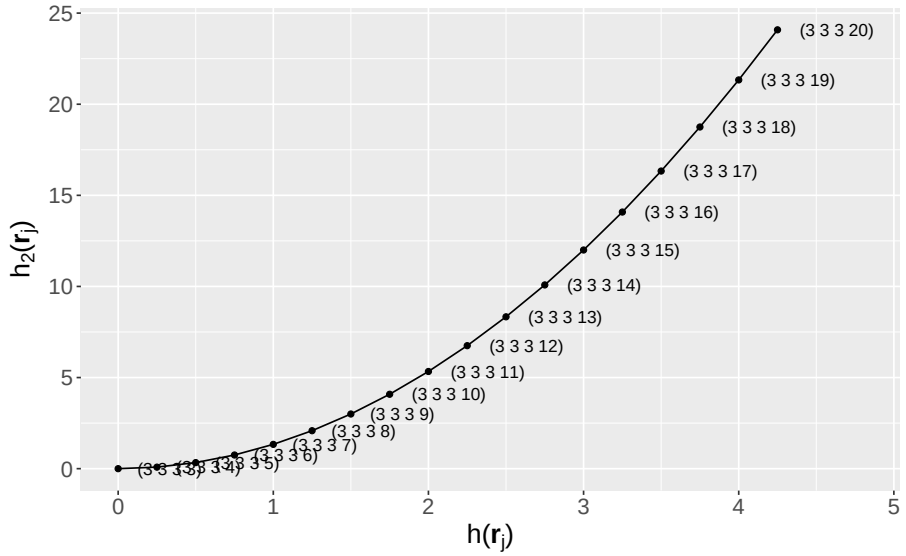


Figure 2.2.: Comparison of $h(\boldsymbol{r}_j)$ and $h_2(\boldsymbol{r}_j)$ for $\boldsymbol{r}_j = (3 \quad 3 \quad 3 \quad i)$ with $i = 3, \ldots, 20$. Measure $h_2(\boldsymbol{r}_j)$ penalizes deviations between $\boldsymbol{r}_j^{-best}$ and r_j^{best} more than $h(\boldsymbol{r}_j)$.

Interestingly, there is a direct link between $h_2(\boldsymbol{r}_j)$ and $h(\boldsymbol{r}_j)$. To see this, reformulate $h(\boldsymbol{r}_j)$:

$$\begin{aligned}
h(\boldsymbol{r}_j) &= \frac{\sum_{k=1}^{m} r_{jk}}{m} - r_j^{best} \\
&= \frac{1}{m}\Big(\sum_{k:r_{jk} \neq r_j^{best}}^{m} r_{jk} + r_j^{best} - m\, r_j^{best} \Big) \\
&= \frac{1}{m}\Big(\sum_{k:r_{jk} \neq r_j^{best}}^{m} r_{jk} - (m-1) r_j^{best} \Big) \\
&= \frac{m-1}{m}\Big(\frac{\sum_{k:r_{jk} \neq r_j^{best}}^{m} r_{jk}}{m-1} - r_j^{best} \Big) \\
&= \frac{m-1}{m}\Big(\overline{\boldsymbol{r}_j^{-best}} - r_j^{best} \Big)
\end{aligned} \tag{2.10}$$

and compare it with $h_2(\boldsymbol{r}_j)$ in Equation (2.9). Thus, the only difference between $h_2(\boldsymbol{r}_j)$ and $h(\boldsymbol{r}_j)$ is the squaring of the term $\big(\overline{\boldsymbol{r}_j^{-best}} - r_j^{best}\big)$. This means that $h_2(\boldsymbol{r}_j)$ penalizes deviations between r_j^{best} and $\boldsymbol{r}_j^{-best}$ more than $h(\boldsymbol{r}_j)$. In Figure 2.2, the difference between the two measures is illustrated by plotting $h(\boldsymbol{r}_j)$ against $h_2(\boldsymbol{r}_j)$ for $\boldsymbol{r}_j = (3 \quad 3 \quad 3 \quad i)$ with $i = 3, \ldots, 20$.

Since the two measures $h(\boldsymbol{r}_j)$ and $h_2(\boldsymbol{r}_j)$ are very similar and because $h(\boldsymbol{r}_j)$ additionally has a graphical interpretation, this thesis will focus on $h(\boldsymbol{r}_j)$. However, $h_2(\boldsymbol{r}_j)$ could be used as well to quantify the data dredging potential of $\boldsymbol{r}_j$, especially if deviations between r_j^{best} and $\boldsymbol{r}_j^{-best}$ are of considerable importance in the respective research question.

2.2.3. Best rank

A drawback of $h(\boldsymbol{r}_j)$ is that it takes the distribution of $\boldsymbol{r}_j^{-best}$ with respect to r_j^{best} into account, but not the actual value of r_j^{best}. For example, consider the ranks $\boldsymbol{r}_j = (1 \quad 7 \quad 8 \quad 10)$ and $\boldsymbol{r}_{j'} = (101 \quad 107 \quad 108 \quad 110)$. Because $h(\boldsymbol{r}_j)$ is equal to 5.5 in both cases, it follows that both results have the same data dredging potential according to $h(\boldsymbol{r}_j)$. However, since $r_j^{best} = 1$ indicates a higher importance of the respective variable than $r_{j'}^{best} = 101$, $\boldsymbol{r}_j = (1 \quad 7 \quad 8 \quad 10)$ is expected to have a higher data dredging potential than $\boldsymbol{r}_{j'} = (101 \quad 107 \quad 108 \quad 110)$.
In the following, two approaches addressing this issue are presented. The underlying ideas are also used for other measures that assess the stability of ranking lists (but do not allow to quantify the data dredging potential as already stated in Section 2.2.1) (Boulesteix and Slawski, 2009).

The first approach consists of modifying $h(\boldsymbol{r}_j)$ such that it adjusts for the value of r_j^{best}. More specifically, $h(\boldsymbol{r}_j)$ is weighted according to the value of r_j^{best}:

$$h(\boldsymbol{r}_j; \alpha) := h(\boldsymbol{r}_j) \cdot (r_j^{best})^{-\alpha}, \tag{2.11}$$

where $\alpha \in [0, \infty)$ is a parameter that can be chosen by the researcher. As in the case of $h(\boldsymbol{r}_j)$, a high value of $h(\boldsymbol{r}_j; \alpha)$ indicates a high data dredging potential. $h(\boldsymbol{r}_j; \alpha)$ can be interpreted as the weighted difference between the average rank of $\boldsymbol{r}_j$ and r_j^{best}. If $r_j^{best} = 1$, $h(\boldsymbol{r}_j; \alpha)$ is equal to $h(\boldsymbol{r}_j)$. This implies that $h(\boldsymbol{r}_j)$ and $h(\boldsymbol{r}_j; \alpha)$ take the same value if $\boldsymbol{r}_j$ corresponds to the worst case ranking result since in this case, r_j^{best} is equal to 1. Consequently, $h(\boldsymbol{r}_j; \alpha)$ has the same range of values as $h(\boldsymbol{r}_j)$, i.e. $h(\boldsymbol{r}_j; \alpha) \in [0, \frac{m-1}{m}(p-1)]$. Parameter α determines the relevance of ranks > 1: the higher α, the smaller $h(\boldsymbol{r}_j; \alpha)$ with increasing r_j^{best}. If $\alpha \to \infty$, ranks > 1 are not considered as relevant and $h(\boldsymbol{r}_j; \infty)$ is equal to zero for all $r_j^{best} > 1$. Thus, according to $h(\boldsymbol{r}_j; \infty)$, a ranking result with $r_j^{best} > 1$ has no data dredging potential. Conversely, if $\alpha = 0$, all ranks are considered as equally relevant and $h(\boldsymbol{r}_j; 0)$ is equal to $h(\boldsymbol{r}_j)$.
Figure 2.3 illustrates the impact of α and r_j^{best} on $h(\boldsymbol{r}_j; \alpha)$ using the example ranking results $\boldsymbol{r}_j = (1 + i \quad 7 + i \quad 8 + i \quad 10 + i)$ with $i = 0, \ldots, 100$. All ranking results yield the same value of $h(\boldsymbol{r}_j)$ but have different values of r_j^{best}. The figure shows that for $\alpha \geqslant 1$, $h(\boldsymbol{r}_j; \alpha)$ is close to zero for ranking results with r_j^{best} larger than approximately 25.
In practice, a good balance between $\alpha = 0$ and $\alpha \to \infty$ might be for instance $\alpha = 0.5$. This has the advantage that $h(\boldsymbol{r}_j; 0.5)$ can be written as

$$h(\boldsymbol{r}_j; 0.5) = \frac{h(\boldsymbol{r}_j)}{\sqrt{r_j^{best}}}. \tag{2.12}$$

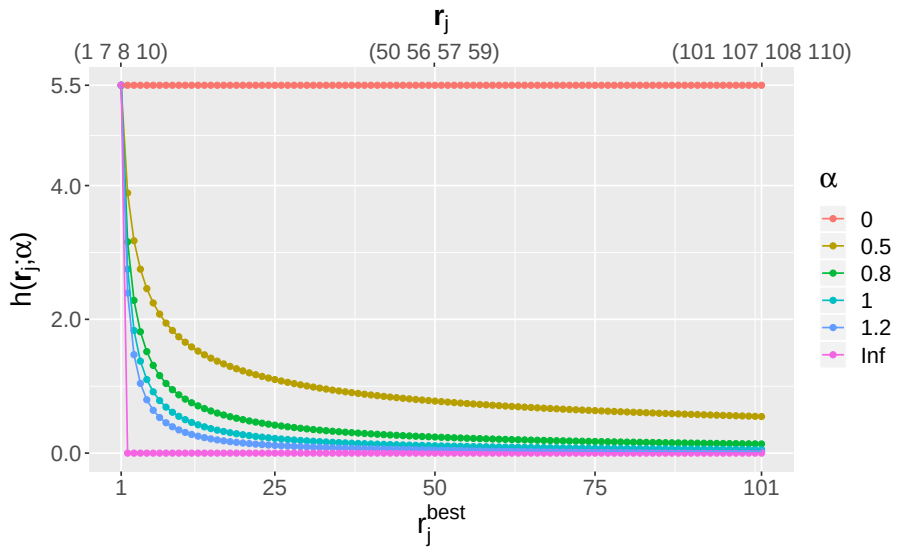


Figure 2.3.: Impact of α on $h(\boldsymbol{r}_j;\alpha)$ for ranking results that yield the same value of $h(\boldsymbol{r}_j)$, $\boldsymbol{r}_j = (1+i \quad 7+i \quad 8+i \quad 10+i)$ with $i = 0,\ldots,100$.

As an example, consider again $\boldsymbol{r}_j = (1 \quad 7 \quad 8 \quad 10)$ and $\boldsymbol{r}_{j'} = (101 \quad 107 \quad 108 \quad 110)$ from the beginning of this section. According to $h(\boldsymbol{r}_j)$, both ranking results have the same data dredging potential ($h(\boldsymbol{r}_j) = h(\boldsymbol{r}_{j'}) = 5.5$). For $\alpha = 0.5$, $h(\boldsymbol{r}_j; 0.5)$ is still equal to 5.5 (because $r_j^{best} = 1$) but $h(\boldsymbol{r}_{j'}; 0.5)$ is approximately equal to 0.55, indicating a lower data dredging potential. To yield a value of $h(\boldsymbol{r}_{j'}; 0.5) = 5.5$ with $r_j^{best} = 101$, the discrepancy between r_j^{best} and $\boldsymbol{r}_j^{-best}$ has to be much higher, for example $\boldsymbol{r}_{j'} = (101 \quad 107 \quad 108 \quad 309)$.

An alternative way to take the actual value of r_j^{best} into account is to choose a cutoff parameter c (e.g $c = 50, 100$) and assign a data dredging potential of zero if $r_j^{best} > c$. This means that

$$\begin{cases} h(\boldsymbol{r}_j), & r_j^{best} \leqslant c, \\ 0, & r_j^{best} > c. \end{cases} \tag{2.13}$$

This approach is much simpler than using $h(\boldsymbol{r}_j;\alpha)$ since the interpretation of $h(\boldsymbol{r}_j)$ remains the same. However, it follows the nothing-or-all principle, which means that all ranking results with $r_j^{best} \leqslant c$ are treated equally with respect to r_j^{best}, whereas all ranking results with $r_j^{best} > c$ are considered as irrelevant (Boulesteix and Slawski, 2009). Section 2.3 will however show that despite its oversimplification, this approach might be useful in combination with $h(\boldsymbol{r}_j;\alpha)$ or even preferable if the overall data dredging potential is assessed.

Both approaches have in common that they require the choice of a parameter (i.e. α or c) that determines the relevance of ranks > 1, which is both advantageous and disadvan-

tageous. On the one hand, this allows the researcher to individually adjust $h(\boldsymbol{r}_j)$ to the respective research question (e.g. up to which rank a variable is relevant enough to be published or a candidate for further research). On the other hand, it raises the problem of finding a suitable parameter value. This might be especially difficult for α because its impact on the relevance of ranks > 1 is less intuitive than that of c. Since there will be inevitably some arbitrariness in the choice of α and c, several parameter values might be considered in practice (Boulesteix and Slawski, 2009).

2.3. Application in practice

In the previous section, a measure that quantifies the data dredging potential of a single ranking result was introduced. In this section, two possible applications in practice are considered, namely rankings with respect to data dredging potential (Section 2.3.1) and the assessment of the overall data dredging potential (Section 2.3.2).

2.3.1. Ranking with respect to data dredging potential

Let again X_{j*} be the researcher's favourite variable and $\boldsymbol{r}_{j*}$ the corresponding ranking result. Based on $h(\boldsymbol{r}_j)$ and its modifications ($h(\boldsymbol{r}_j;\alpha)$ or cutoff c), it is now possible to quantify the data dredging potential of $\boldsymbol{r}_{j*}$. Additionally, it might be interesting to compare the data dredging potential of $\boldsymbol{r}_{j*}$ with the data dredging potential of the other $p-1$ ranking results.
A possible way to do this is to rank the variables $X_1, \ldots, X_p$ according to the data dredging potential of $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_p$ (not to be confused with the original rankings where the variables were ordered with respect to their relevance for Y). This approach allows the researcher to assess the rank of X_{j*} with respect to the data dredging potential of $\boldsymbol{r}_{j*}$ and to check if $\boldsymbol{r}_{j*}$ is among the ranking results with the highest data dredging potential.
As already stated in the previous section, there are two approaches that take the actual value of r_j^{best} into account. To rank the variables according to their data dredging potential, a combination of both approaches might be an appropriate strategy: By choosing a cutoff value c, all ranking results that are definitely not relevant are sorted out. The α parameter in $h(\boldsymbol{r}_j;\alpha)$ then adjusts the data dredging potential of the remaining ranking results with respect to their best rank.

2.3.2. Assessment of overall data dredging potential

Even if the researcher has no specific favourite variable, it might be useful to quantify the overall data dredging potential of the ranking results. This gives the researcher an overview about how unstable the results actually are with respect to data dredging.
A graphical way to get a first impression of the overall data dredging potential without

using $h(\boldsymbol{r}_j)$ is to generate boxplots that show the distribution of the ranks in $\boldsymbol{r}_j$ for each variable. Another option would be to plot the number of variables with $r_j^{best} \leqslant c$ against $c = 1, \ldots, c_{max}$ (e.g. $c_{max} = 100$). If there is no data dredging potential in the ranking results with $r_j^{best} \leqslant c$ (i.e. all ranks in $\boldsymbol{r}_j$ are equal), the number of variables with $r_j^{best} \leqslant c$ is equal to c.

Another approach to assess the overall data dredging potential is to use an aggregated version of $h(\boldsymbol{r}_j)$ or $h(\boldsymbol{r}_j; \alpha)$. For this purpose, define the mean of $h(\boldsymbol{r}_j)$ and $h(\boldsymbol{r}_j; \alpha)$ over all ranking results with $r_j^{best} \leqslant c$ as

$$H(c) := \frac{1}{|\mathcal{E}_c|} \sum_{j \in \mathcal{E}_c} h(\boldsymbol{r}_j), \tag{2.14}$$

$$H(c; \alpha) := \frac{1}{|\mathcal{E}_c|} \sum_{j \in \mathcal{E}_c} h(\boldsymbol{r}_j; \alpha), \tag{2.15}$$

where c is again a cutoff parameter and $\mathcal{E}_c = \{j : r_j^{best} \leqslant c\}$ is the set of variables with $r_j^{best} \leqslant c$. Equations (2.14) and (2.15) show that both $H(c)$ and $H(c; \alpha)$ take r_j^{best} into account by considering only ranking results with $r_j^{best} \leqslant c$. $H(c)$ can be interpreted as the mean difference between and $\bar{\boldsymbol{r}}_j$ and r_j^{best} over all variables with j in $\mathcal{E}_c$. $H(c; \alpha)$ additionally uses $h(\boldsymbol{r}_j; \alpha)$ instead of $h(\boldsymbol{r}_j)$ which results in a less intuitive interpretation: $H(c; \alpha)$ is the weighted mean difference between $\bar{\boldsymbol{r}}_j$ and r_j^{best} over all variables with j in $\mathcal{E}_c$ (with weights according to r_j^{best}). Actually, it may not be necessary to use $H(c; \alpha)$ if $H(c)$ is considered for more than one cutoff value. For example, $H(c)$ can be plotted against $c \in [1, c_{max}]$ (e.g. $c_{max} = 100$), which leverages the nothing-or-all principle.

When using $H(c; \alpha)$ or $H(c)$, it has to be noted that in general, r_j^{best} is not obtained by the same method for each $j \in \mathcal{E}_c$. However, the definition of data dredging given in the beginning of this chapter implies that only one method is reported (namely the method that yields the best-looking result). This means that because r_j^{best} is not achieved by one single method for all $j \in \mathcal{E}_c$, the "full" data dredging potential can in general not be exploited for all variables at once.

Besides the assessment of overall stability with respect to data dredging, $H(c)$ and $H(c; \alpha)$ can be used to identify the method that yields on average the highest increase in data dredging potential if it is added to the set of ranking methods. For this purpose, let $H^{-k}(c)$ be the value of $H(c)$ that arises when the rankings are generated without considering method k ($k = 1, \ldots, m$). The method that yields on average the highest increase in data dredging potential is the method that yields the smallest value of $H^{-k}(c)$ and should be used with caution (the same procedure can be applied to $H(c; \alpha)$). However, it has to be noted that a method that increases the instability of results does not necessarily yield incorrect ranking results (Boulesteix and Slawski, 2009).

To summarize, there are three steps that should be performed to assess the overall data dredging potential of all ranking results:

1. Consider the distribution of ranks in each $\boldsymbol{r}_j$ and the number of variables with $r_j^{best} \leqslant c$ for $c = 1, \ldots, c_{max}$.
2. Quantify the overall data dredging potential using $H(c)/H(c;\alpha)$, e.g. by plotting $H(c)$ against $c = 1, \ldots, c_{max}$.
3. Optional: Identify the method that yields on average the highest increase in data dredging potential.

These steps can also be performed to compare the overall data dredging potential resulting from different data sets or different numbers/types of ranking methods.

The following chapters illustrate the quantitative assessment of the data dredging potential based on simulated (Chapter 3) and real data (Chapter 4 and 5). To rank the variables with respect to their data dredging potential, a combination of $h(\boldsymbol{r}_j;\alpha)$ and c will be used. For the assessment the overall data dredging potential, only $H(c)$ will be considered because of its more intuitive interpretation.

3. Simulation study

In this chapter, the framework for the quantification of data dredging potential in ranking results is applied to simulated gene expression data in a two-group setting. In Section 3.1 and 3.2, the data generation process and the considered ranking methods are described. In Section 3.3, the data dredging potential is quantified and compared across different parameter settings.

3.1. Simulation design

To study the impact of different parameter settings on the data dredging potential, 27 data sets are generated. They differ with respect to the number of variables, the number of observations and the correlation structure. Each data set $\mathcal{D}(\boldsymbol{x}, \boldsymbol{y})$ consists of a matrix $\boldsymbol{x} = (x_{ij})_{\substack{i=1,\dots,n \\ j=1,\dots,p}}$ and a vector $\boldsymbol{y} = (y_1, \dots, y_n)^\top$. The matrix $\boldsymbol{x}$ contains n independent observations of the variables $X_1, \dots, X_p$, which represent log-transformed expression levels of p genes. In this simulation, a two-group scenario is considered. This means that vector $\boldsymbol{y}$ stores n independent observations of a binary variable Y. For example, Y could be an outcome of interest or an experimental condition fixed by design. In the following, the two possible values of Y are referred to as *group 1* and *group 2*. Both groups are of equal size, i.e. $n_1 = n_2 = \frac{n}{2}$. (Boulesteix and Slawski, 2009; Wu, 2005)
In each data set, the first 100 genes are *differentially expressed* (*DE*). A gene X_j is truly differentially expressed if its expected values of expression are not equal between the two groups of interest, i.e. $\mu_{j1} \neq \mu_{j2}$ (Jeanmougin et al., 2010). In this case, X_j is related to Y and therefore relevant. Conversely, a gene is referred to as *non-differentially expressed* (*noDE*) if its expected values of expression are equal between the two groups of interest, i.e. $\mu_{j1} = \mu_{j2}$ (Jeanmougin et al., 2010). The remaining $p - 100$ genes in each data set are non-differentially expressed.
To study the impact of n and p on the data dredging potential, the number of observations varies with $n \in \{20, 40, 60\}$ and the number of genes varies with $p \in \{2000, 5000, 10000\}$. Moreover, three different scenarios for the correlation structure are considered, which will be described in the following.

Independent genes

The approach for generating independent gene expression levels is based on Irigoien and Arenas (2018). As stated above, the expected means of a DE gene are not equal between the two groups. For each observation $i = 1, \ldots, n$, the DE genes $j = 1, \ldots, 100$ are simulated from the following univariate normal distribution:

$$\begin{cases} N(1, \sigma_{j1}), & \text{if observation i is in group 1,} \\ N(1 + \delta_j, \sigma_{j2}), & \text{if observation i is in group 2,} \end{cases} \tag{3.1}$$

where δ_j is drawn from a uniform distribution on $[0.3, 2]$ and the standard deviations σ_{j1} and σ_{j2} are randomly selected among $\{1.3, 1.5, 1.7, 2.0\}$. This implies that each gene has not necessarily the same variance in group 1 and 2. Moreover, since the difference in the expected means (i.e. δ_j) differs for every $j = 1, \ldots, 100$, the DE genes are not equally relevant for Y.
For the remaining noDE genes, i.e. for $j = 101, \ldots, p$, each observation $i = 1, \ldots, n$ is drawn from

$$\begin{cases} N(1, \sigma_{j1}), & \text{if observation i is in group 1,} \\ N(1, \sigma_{j2}), & \text{if observation i is in group 2.} \end{cases} \tag{3.2}$$

where σ_{j1} and σ_{j2} are again randomly selected among $\{1.3, 1.5, 1.7, 2.0\}$.

Correlated genes

In the scenario of correlated genes, each observation $i = 1, \ldots, n$ is drawn from a multivariate normal distribution:

$$\begin{cases} N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), & \text{if patient i is in group 1,} \\ N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), & \text{if patient i is in group 2.} \end{cases} \tag{3.3}$$

The mean vectors are defined as

$$\underset{p\times 1}{\boldsymbol{\mu}_1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad \underset{p\times 1}{\boldsymbol{\mu}_2} = \begin{pmatrix} 1 + \delta_1 \\ \vdots \\ 1 + \delta_{100} \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

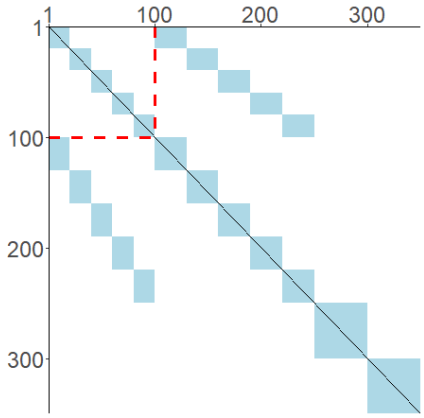


Figure 3.1.: Block-structured correlation matrix for genes $j = 1, \ldots, 350$; the first 100 genes are differentially expressed. Correlations that are equal to ρ are highlighted in blue.

The values $\delta_1, \ldots, \delta_{100}$ are drawn from a uniform distribution on $[0.3, 2]$. The variance-covariance matrices are defined as

$$\underset{p\times p}{\boldsymbol{\Sigma}_1} = \text{diag}(\sigma_{11}, \ldots, \sigma_{p1})\, \boldsymbol{Cor}\, \text{diag}(\sigma_{11}, \ldots, \sigma_{p1}), \tag{3.4}$$

$$\underset{p\times p}{\boldsymbol{\Sigma}_2} = \text{diag}(\sigma_{12}, \ldots, \sigma_{p2})\, \boldsymbol{Cor}\, \text{diag}(\sigma_{12}, \ldots, \sigma_{p2}), \tag{3.5}$$

where $\underset{p\times p}{\boldsymbol{Cor}}$ denotes the correlation matrix and $\text{diag}(\sigma_{11}, \ldots, \sigma_{p1})$ and $\text{diag}(\sigma_{12}, \ldots, \sigma_{p2})$ are diagonal matrices. As in the case of independent genes, σ_{j1} and σ_{j2} are randomly selected among $\{1.3, 1.5, 1.7, 2.0\}$ for each gene X_j. The structure of the correlation matrix $\boldsymbol{Cor}$ is based on Korn et al. (2004). It is generated by dividing the genes into blocks of size 50. This means that depending on the number of genes ($p = 2000, 5000, 10000$), there are 40, 100 or 200 blocks. Moreover, the first five blocks contain 20 DE genes each. The pairwise correlation between genes within a block is ρ, whereas the correlation between the blocks is 0. The correlation of a gene with itself is equal to 1. Figure 3.1 displays the correlation matrix for the first 350 genes in each data set. In the simulation, $\rho = 0.4$ and $\rho = 0.8$ are considered. Note that $\rho = 0$ corresponds to the scenario of independent genes described above.

Overall, the combination of $n \in \{20, 40, 60\}$, $p \in \{2000, 5000, 10000\}$ and $\rho \in \{0, 0.4, 0.8\}$ yields 27 data sets. In each data set, the DE genes are labelled as *DE1,...,DE100* and the noDE genes as *noDE101,...,noDEp*.

3.2. Ranking methods

The genes in each data set are ranked by seven methods that differ with respect to their ranking criterion. An overview of all used methods can be found in Table 3.1.

Table 3.1.: Overview of ranking methods used in the simulation.

k	method
1	Fold-change
2	T-statistic
3	Permutation test
4	SAM
5	Limma
6	Welch's t-statistc
7	Wilcoxon statistic

The considered methods are univariate, which means that in contrast to multivariate methods, they evaluate the relevance of each gene independently from the other genes (Dessì et al., 2013). In principle, all methods are based on testing the null hypothesis that the expected values of expression for a given gene X_j are equal between the two groups, i.e. $H_0 : \mu_{j1} = \mu_{j2}$ vs. $H_1 : \mu_{j1} \neq \mu_{j2}$ for each $j = 1, \ldots, p$ (Jeanmougin et al., 2010). The ranking criteria are defined as the test statistic or p-value of a suitable test (e.g. the two-sample t-test). The higher the absolute value of the test statistic/the smaller the p-value, the smaller the assigned rank.
Ideally, none of the ranking methods should assign a rank > 100 to one of the DE genes in the simulated data sets. Moreover, the most differentially expressed gene (i.e. the gene with the highest value of δ_j) should be assigned a rank of 1 by all methods.

The rest of this section briefly reviews the considered ranking methods, which are implemented in the Bioconductor add-on package `GeneSelector` (Slawski and Boulesteix, 2018). To describe the ranking methods, let $\bar{x}_{j1}$ and $\bar{x}_{j2}$ denote the sample means of X_j for group 1 and 2, respectively. Moreover, the pooled sample variance of X_j is given by

$$s_j^2 = \frac{(n_1 - 1)s_{j1}^2 + (n_2 - 1)s_{j2}^2}{n_1 + n_2 - 2}, \tag{3.6}$$

where s_{j1}^2 and s_{j2}^2 are the group specific sample variances with $s_{jg}^2 = \frac{1}{n_g - 1} \sum_{i:y_i=g} (x_{ij} - \bar{x}_{jg})$ for $g \in \{1, 2\}$. (Jeanmougin et al., 2010)
The considered methods are the following:

- **Fold-change**: Assuming that the expression levels are log-transformed, the ranking criterion is simply the absolute value of the mean difference between the two groups

(Boulesteix and Slawski, 2009). For gene X_j, this is equal to

$$|FC_j| = |\bar{x}_{j1} - \bar{x}_{j2}|. \tag{3.7}$$

Note this procedure is very naive since it only considers the difference in means without taking variances into account (Slawski and Boulesteix, 2018).

- **T-statistic**: The ranking criterion is the absolute value of the test statistic resulting from the ordinary t-test for two groups. For gene X_j, the test statistic is defined as

$$t_j^{Tstat} = \frac{\bar{x}_{j1} - \bar{x}_{j2}}{s_j \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.8}$$

(Jeanmougin et al., 2010).

- **Permutation test**: This method uses the permutation p-values as ranking criterion. The underlying idea of a permutation test is to estimate the distribution of a test statistic under the null hypothesis and compare it to the actual value of the test statistic. In this simulation, the considered test statistic is t_j^{Tstat} and the null hypothesis is $H_0 : \mu_{j1} = \mu_{j2}$. The distribution of H_0 is estimated by generating B (e.g. $B = 100$) random permutations of the values in $\boldsymbol{y}$ and calculating the test statistic $t_{j,b}^{Tstat}$ for each permutation $b = 1, \ldots, B$. The permutation p-value of X_j is then defined as

$$p_j^{Perm} = \frac{\sum_{b=1}^{B} \mathbb{1}\big(|t_j^{Tstat}| < |t_{j,b}^{Tstat}|\big)}{B}. \tag{3.9}$$

If two or more variables yield the same permutation p-values, $|t_j^{Tstat}|$ is used as second-order ranking criterion. (Dudoit et al., 2003; Slawski and Boulesteix, 2018)

- **SAM (Significance Analysis of Microarrays)**: The ranking criterion is the absolute value of the SAM statistic t_j^{SAM}, which is a modified version of the t-statistic. It stabilises t_j^{Tstat} by adding a small positive constant s_c to the denominator:

$$t_j^{SAM} = \frac{\bar{x}_{j1} - \bar{x}_{j2}}{s_c + s_j \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.10}$$

(Tusher et al., 2001). The value of the constant is only computed once for all genes and is chosen to minimize the coefficient of variation of the test statistic, for details see Chu et al. (2002).

- **Limma (Linear Models for Microarray Data)**: This method uses the absolute value of the Limma statistic as ranking criterion. The Limma statistic represents a hybrid classical/Bayes approach in which the posterior variance $s_j^{2\,Limma}$ is substi-

tuted into the classical t-statistic:

$$t_j^{Limma} = \frac{\bar{x}_{j1} - \bar{x}_{j2}}{s_j^{Limma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{3.11}$$

where ${s_j^2}^{Limma}$ is a weighted combination of an estimate obtained from the prior distribution (s_0^2) and the pooled sample variance (s_j^2):

$${s_j^2}^{Limma} = \frac{d_0 s_0^2 + d_j s_j^2}{d_0 + d_j}, \tag{3.12}$$

where d_0 and d_j are prior and empirical degrees of freedom, respectively. Using a prior distribution has the effect of borrowing information from the ensemble of genes for inference about each individual gene. It shrinks the observed variances towards the prior values. (Smyth, 2004; Jeanmougin et al., 2010)

- **Welch's t-statistic** : The ranking criterion is the absolute value of the test statistic resulting from the t-test for two groups with unequal variances:

$$t_j^{WelchT} = \frac{\bar{x}_{j1} - \bar{x}_{j2}}{\sqrt{\frac{s_{j1}^2}{n_1} + \frac{s_{j2}^2}{n_2}}} \tag{3.13}$$

(Jeanmougin et al., 2010).

- **Wilcoxon statistic**: This method uses the absolute value of the Wilcoxon statistic as ranking criterion. The Wilcoxon statistic of gene X_j is defined as

$$W_j^{Wilcox} = R_j - \frac{n_1(n_1+1)}{2}, \tag{3.14}$$

where R_j is the rank sum. R_j is calculated by combining the expression levels of group 1 and 2 for gene X_j, listing them in rank order and taking the sum of all ranks that belong to the group with more observations (or simply group 1 if $n_1 = n_2$). (Jeanmougin et al., 2010; Slawski and Boulesteix, 2018)

3.3. Results

In this section, the ranking results are analysed with respect to their data dredging potential. This includes comparisons across the parameter settings as well as comparisons between DE genes and noDE genes.
As stated in Chapter 2, $h(\boldsymbol{r}_j)$ and its modifications ($h(\boldsymbol{r}_j;\alpha)$ and cutoff parameter c) do not only allow to quantify the data dredging potential of a single ranking result but also

Table 3.2.: Top-10 lists of ranking results with the highest data dredging potential resulting from $c = 100$ and $\alpha = \{0, 0.5, 0.8\}$ for a simulated data set with $p = 2000$, $n = 40$, $\rho = 0$. Empty cells denote that the ranking result is not among the top-10. For each ranking result, r_j^{best} is highlighted in red.

gene	$\boldsymbol{r}_j$	$h(\boldsymbol{r}_j;\alpha)$			rank (w.r.t. $h(\boldsymbol{r}_j;\alpha)$)		
		$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$
noDE390	(482 230 167 353 366 230 61)	208.86	26.74	7.79	1	1	1
noDE1510	(72 312 345 182 187 312 436)	191.71	22.59	6.26	2	2	3
noDE744	(87 262 249 158 162 262 406)	139.57	14.96	3.92	3	6	
noDE1511	(77 264 250 153 156 264 349)	139.14	15.86	4.31	4	4	7
noDE1951	(99 257 295 163 167 257 398)	134.71	13.54	3.41	5	9	
noDE309	(237 203 224 214 214 203 69)	125.86	15.15	4.25	6	5	9
noDE1293	(348 102 47 202 226 102 115)	116.14	16.94	5.34	7	3	4
noDE1165	(84 241 208 145 148 241 328)	115.29	12.58	3.33	8		
noDE1613	(226 193 75 198 200 193 221)	111.57	12.88	3.53	9		
noDE1182	(279 177 285 219 219 177 97)	110.57	11.23	2.85	10		
noDE1890	(264 111 49 180 183 111 129)	97.71	13.96	4.34		7	6
noDE1936	(49 164 196 95 96 164 256)	96.71	13.82	4.30		8	8
DE3	(41 135 173 76 77 135 245)	85.00	13.27	4.36		10	5
DE4	(1 10 10 8 8 10 12)	7.43	7.43	7.43			2
noDE1457	(155 16 15 61 66 16 20)	34.86	9.00	3.99			10

to rank the variables according to their data dredging potential and to assess the overall data dredging potential. Both applications are considered in the following.

Rankings with respect to data dredging potential

First, the genes are ranked with respect to their data dredging potential. For this purpose, the data dredging potential of the ranking results $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_p$ in each data set is assessed by a combination of $h(\boldsymbol{r}_j;\alpha)$ and a cutoff value of $c = 100$. To illustrate the impact of α on the data dredging potential, several values for α are considered. As stated in Chapter 2, $h(\boldsymbol{r}_j;\alpha)$ is close to zero for most ranking results if $\alpha \geqslant 1$. For this reason, α varies with $\alpha \in \{0, 0.5, 0.8\}$. However, it has to be noted that the choice of α is not based on an objective criterion and that other values for α could be used as well.

After assessing the data dredging potential of each ranking result, the genes are ordered such that a small rank of gene X_j indicates a high data dredging potential of $\boldsymbol{r}_j$ (note that in the original rankings, a small value of X_j indicates a high relevance for Y). Since three values for α are considered, there are three different rankings with respect to data dredging potential for each data set.

Table 3.2 displays the top-10 lists of ranking results with the highest data dredging potential for a simulated data set with 2000 uncorrelated genes and 40 observations (i.e. $\rho = 0$, $p = 2000$ and $n = 40$). It shows that for each α value, gene *noDE390* has the highest data dredging potential. For $\alpha = 0$, $h(\boldsymbol{r}_{390};0)$ takes a value of 208.86, which means that the difference between the average rank and the best rank of this gene is equal to 208.86. For $\alpha = 0.5$ and $\alpha = 0.8$, the value of $h(\boldsymbol{r}_{390};\alpha)$ is reduced to 26.74 and 7.79, respectively.

Apart from gene *noDE390*, the three top-10 lists in Table 3.2 differ depending on the value of α. As explained in Chapter 2, a parameter value of $\alpha = 0$ only considers the distribution of $\boldsymbol{r}_j^{-best}$ with respect to r_j^{best}, whereas a value of $\alpha > 0$ additionally takes the actual value of r_j^{best} into account. This is reflected in Table 3.2: For $\alpha = 0$, the top-10 list only includes ranking results with best ranks larger than 49 and a high difference between average and best rank; the actual value of r_j^{best} is not of relevance (except for $r_j^{best} \leqslant 100$). Conversely, for $\alpha \in \{0.5, 0.8\}$, the top-10 ranking results tend to have smaller values of r_j^{best} and a smaller range.
The impact of α on the data dredging potential is also illustrated in Figure 3.2, which displays all ranking results with $r_j^{best} \leqslant 100$ for the same simulated data set as shown in Table 3.2. The figure consists of three panels that highlight the top-10 ranking results with the highest data dredging potential for each $\alpha \in \{0, 0.5, 0.8\}$. To simplify the comparison between the panels, r_j^{best} is plotted against $h(\boldsymbol{r}_j; 0)$ in each panel.
Both Table 3.2 and Figure 3.2 show that for $\alpha = 0$, all genes in the top-10 list are noDE genes. As can be seen from Figure 3.2, this is because noDE genes tend to have a higher variability in ranking results and higher values of r_j^{best} than DE genes. If the actual value of r_j^{best} is additionally taken into account (i.e. for $\alpha > 0$), the 10-top list also includes DE genes.
When looking at the best rank of each ranking result in Table 3.2, it is striking that r_j^{best} is exclusively assigned by the ranking methods using the Wilcoxon statistic, the permutation test or the fold-change criterion. Consequently, the data dredging potential of the ranking results in Table 3.2 would be considerably smaller if these three methods were omitted. Ensuing results being depicted in this section will show that this does not only apply to the top-10 list but also to other ranking results.
Overall, the results shown for the data set with $p = 2000$, $n = 40$ and $\rho = 0$ are in line with the results of the other simulated data sets. Corresponding tables and figures can be found in the electronic appendix.

Overall data dredging potential

This section aims to provide an overview about how unstable the ranking results of each simulated data set are with respect to data dredging. As explained in Chapter 2, there are three steps that should be performed to assess the overall data dredging potential. In the following, this procedure is applied to the simulated data sets.

Step 1. Consider the distribution of ranks and the number of variables with $r_j^{best} \leqslant c$

To get a first impression of the overall data dredging potential, the distribution of ranks and the number of variables with $r_j^{best} \leqslant c$ can be considered. As an example, Figure 3.3 shows the number of variables with $r_j^{best} \leqslant c$, $c \in [1, 100]$, for six simulated data sets with $\rho = 0$, $p \in \{2000, 10000\}$ and $n \in \{20, 40, 60\}$. If all ranking methods yield the same rank

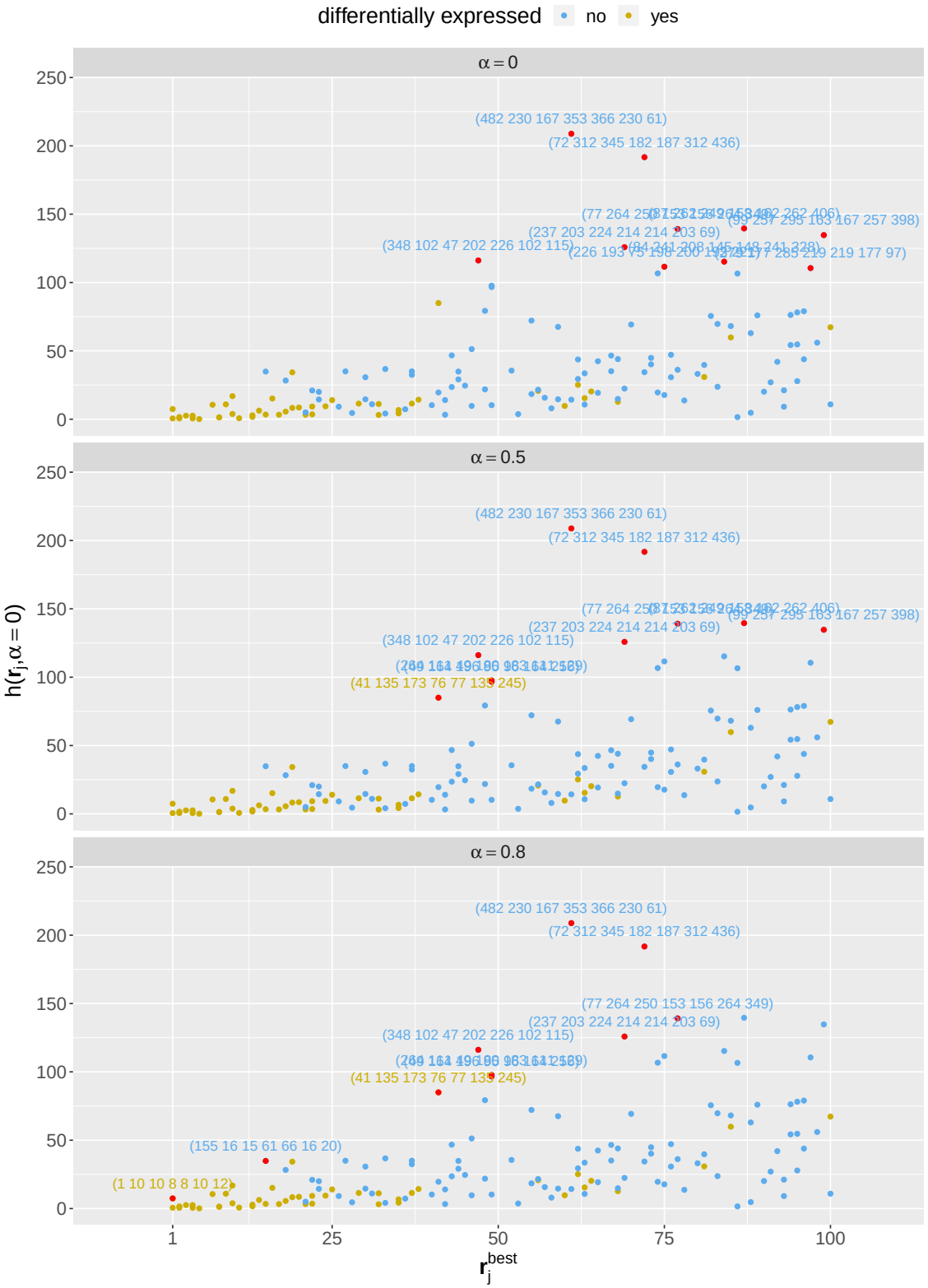


Figure 3.2.: Data dredging potential of ranking results with $r_j^{best} \leqslant 100$ for a simulated data set with $p = 2000$, $n = 40$, $\rho = 0$. In each panel, the top-10 ranking results with the highest data dredging potential for $\alpha \in \{0, 0.5, 0.8\}$ are highlighted in red.

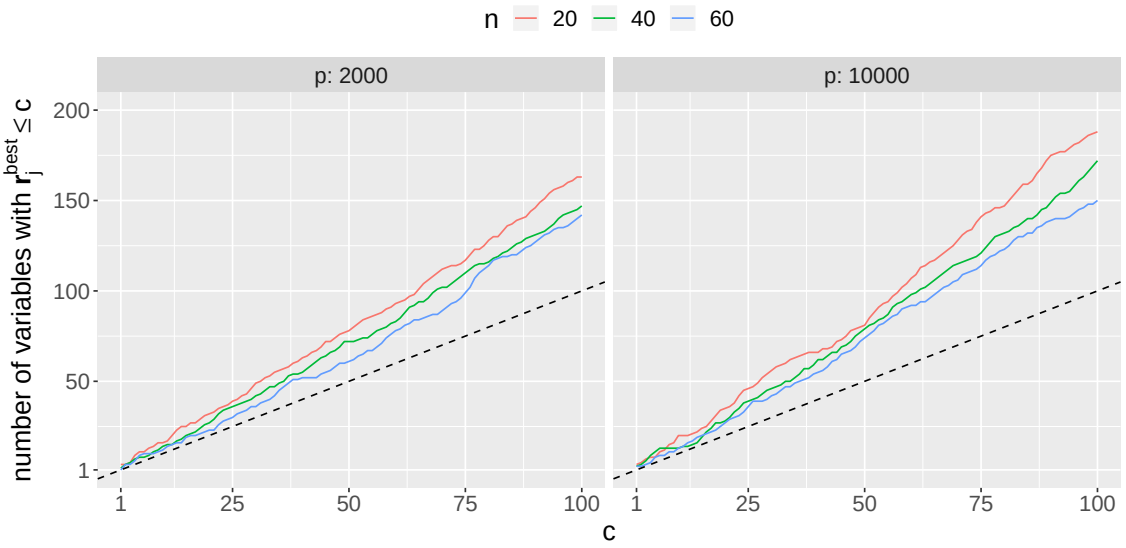


Figure 3.3.: Number of variables with $r_j^{best} \leqslant c$, $c \in [1, 100]$, for simulated data sets with $p \in \{2000, 10000\}$ and $n \in \{20, 40, 60\}$ and $\rho = 0$. If all methods yield the same rank for each variable, the number of variables with $r_j^{best} \leqslant c$ is equal to c (dashed line).

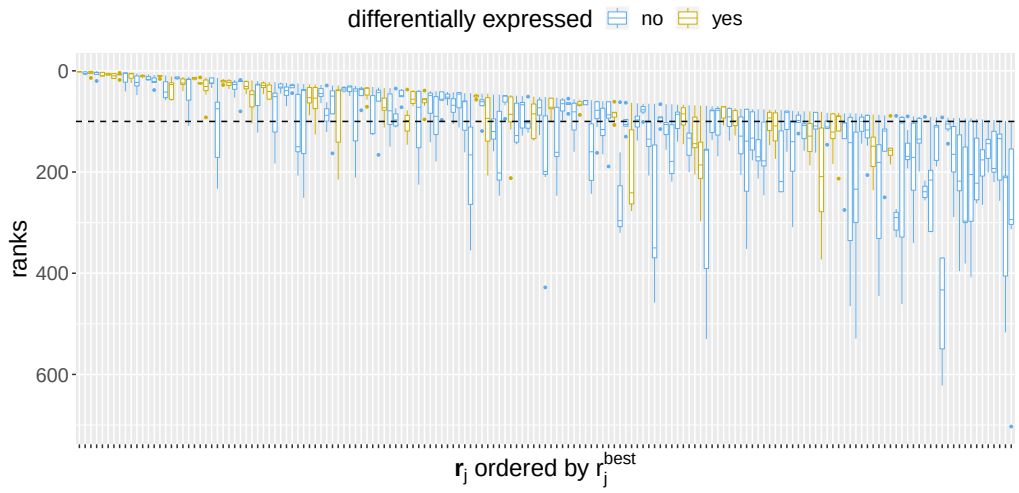
for each variable (i.e. if the ranking results have no data dredging potential), the number of variables with $r_j^{best} \leqslant c$ is equal to c for each $c \in [1, p]$. On contrary, if the number of variables with $r_j^{best} \leqslant c$ is much higher than c, this indicates a high data dredging potential since there are many variables with a small best rank.

Figure 3.3 shows that the higher n and the smaller p, the higher the number of variables with $r_j^{best} \leqslant c$ for $c \in [1, 100]$. This indicates that the overall data dredging potential increases with increasing p and decreasing n.

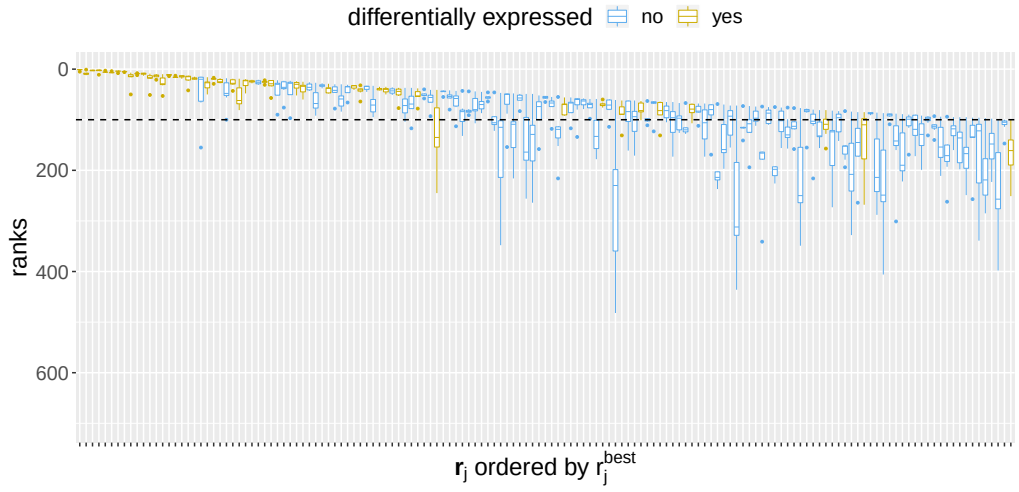
A second way to get a general idea of the overall data dredging potential is illustrated in Figure 3.4 and 3.5. Both figures contain boxplots showing the distribution of each ranking result with $r_j^{best} \leqslant 100$ for the same six data sets as presented in Figure 3.3.

When looking at the figures that display the results for $p = 2000$ and $n \in \{20, 40, 60\}$ (Figures 3.4a-3.4c), it can be observed that the variability within the ranking results $\boldsymbol{r}_j$ increases if the number of observations decreases. Moreover, the figures show that the smaller the number of observations, the smaller the value of r_j^{best} for noDE genes. As stated above, all noDE genes should ideally have ranks > 100. The results for $p = 10000$ (Figures 3.5a-3.5c) show similar tendencies regarding the impact of n. Furthermore, Figures 3.4 and 3.5 disclose that the variability within the ranking results increases with the number of variables. This applies in particular to the data sets with $n = 20$.

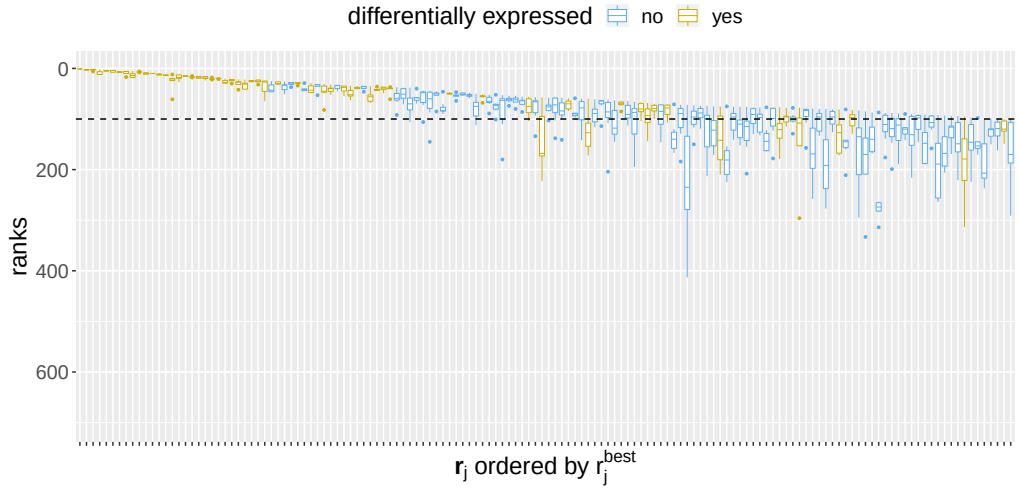
Finally, Figure 3.4 and 3.5 show that the variability within the ranking results of noDE genes is generally higher than in DE genes. This confirms the findings from Table 3.2 and Figure 3.2.



(a) $n = 20$

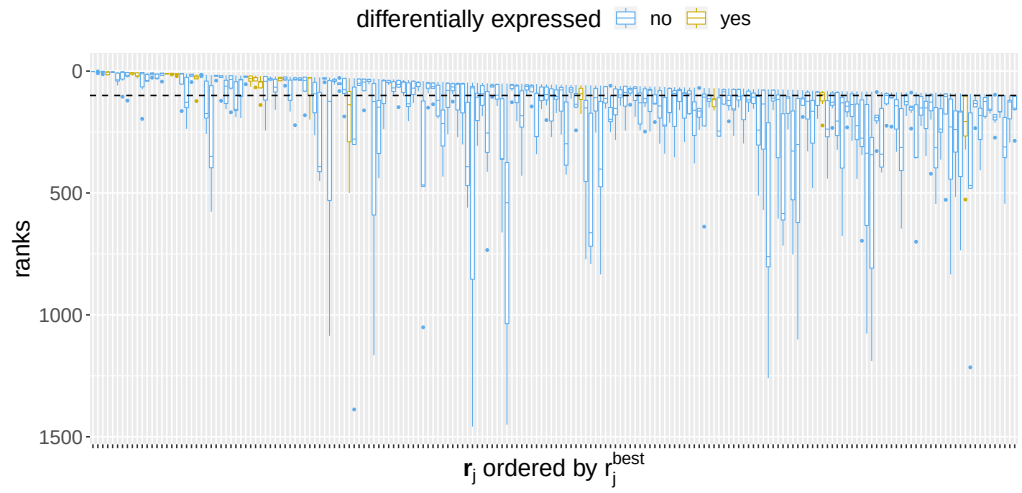


(b) $n = 40$

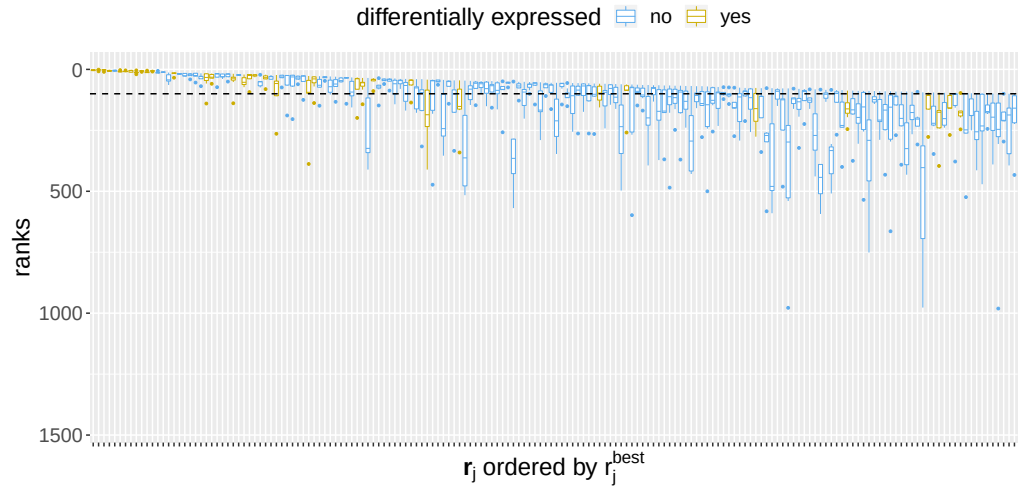


(c) $n = 60$

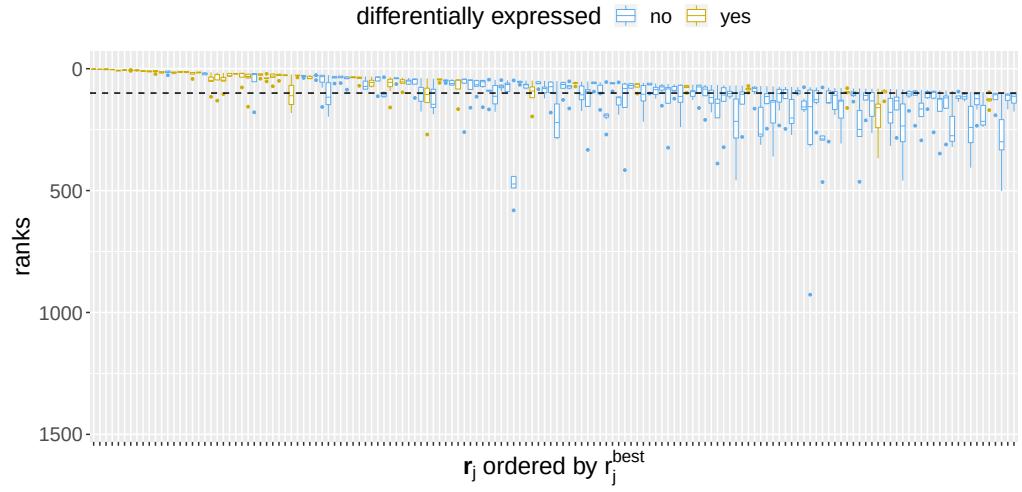
Figure 3.4.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0$, $p = 2000$ and $n \in \{20, 40, 60\}$.



(a) $n = 20$



(b) $n = 40$



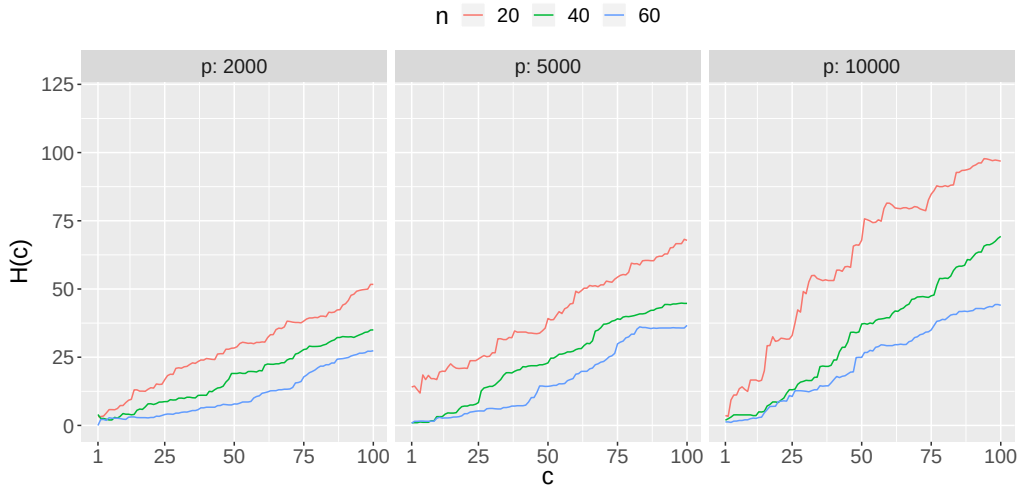
(c) $n = 60$

Figure 3.5.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0$, $p = 10000$ and $n \in \{20, 40, 60\}$.

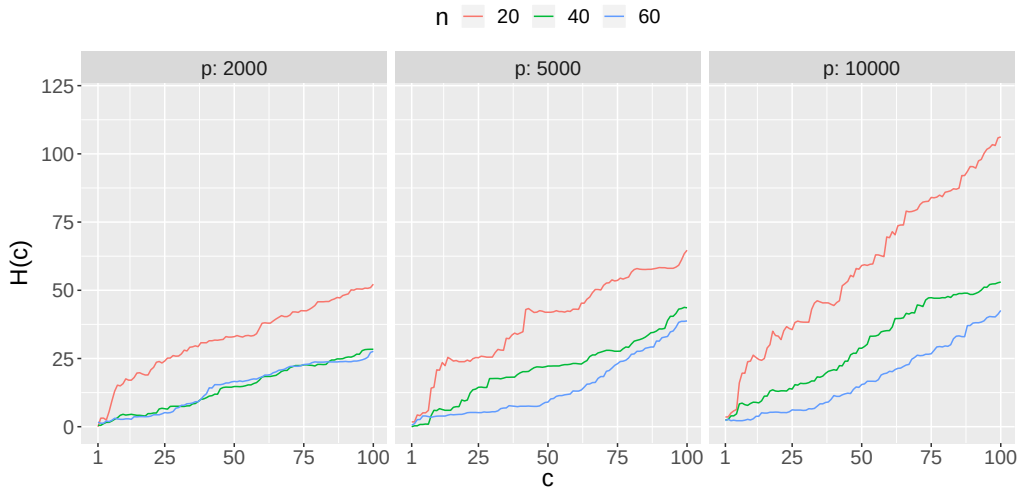
The other simulated data sets provide similar results and point to the same conclusions. The correlation structure between the genes (i.e. parameter ρ) does not appear to affect the variability within the ranking results. Corresponding figures are provided in the appendix (Figures A.1 and A.2 - A.8).

Step 2. Quantify the overall data dredging potential using $H(c)$

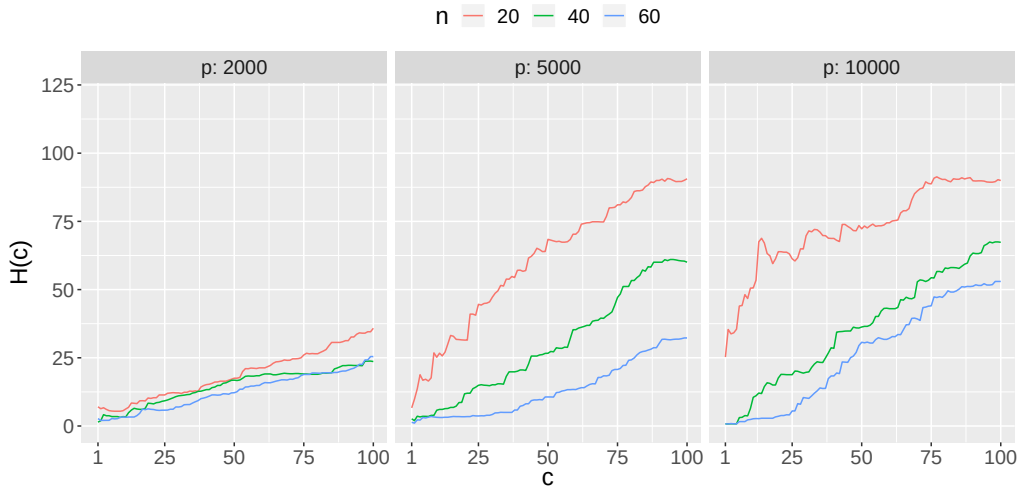
As a second step, the overall data dredging potential is assessed quantitatively. In Figure 3.6, $H(c)$ is plotted against $c \in [1, 100]$ for all 27 simulated data sets. For example, the second panel of Figure 3.6b displays $H(c)$ for three data sets with $\rho = 0.4$, $p = 5000$ and $n \in \{20, 40, 60\}$. For $n = 60$ and $c = 75$, the value of $H(c)$ is equal to 25, whereas for $n = 20$ it is larger than 50. Consequently, the mean difference between the average and the best rank for all ranking results with $r_j^{best} \leqslant 75$ is equal to 25 for $n = 60$ and larger than 50 for $n = 20$.
Figure 3.6 shows that in each data set, the value of $H(c)$ tends to increase with the cutoff value c. This implies that the higher the value of r_j^{best}, the larger the difference between the average and best rank of $\boldsymbol{r}_j$. Moreover, it can be observed that the overall data dredging potential increases with decreasing number of observations (for fixed values of c, ρ and p). On contrary, the number of variables p seems to increase the overall data dredging potential. However, this mainly applies to data sets with $n = 20$.
To ensure that the observed tendencies regarding the impact of c, p and n on the data dredging potential are not due to random fluctuations, 20 simulated data sets were additionally generated for each combination of n, p and ρ. Figure 3.7 displays the resulting mean values of $H(c)$ for each parameter setting. The figure confirms that $H(c)$ increases with increasing c and decreasing number of observations. As already observed in Figure 3.6, the number of variables affects the data dredging potential mainly for $n = 20$. Regarding the impact of the correlation between the variables, there is no noticeable difference between the figures for $\rho = 0$, $\rho = 0.4$ and $\rho = 0.8$.
To summarize, the overall data dredging potential depends on the cutoff parameter c, the number of observations n and the number of variables p. The results are in line with the results of Step 1, which were shown in Figures 3.3, 3.4 and 3.5.
When comparing different $H(c)$ values, the question may arise whether there is a threshold for $H(c)$ (or $h(\boldsymbol{r}_j)$ if a single $\boldsymbol{r}_j$ is considered) indicating that the ranking results are unproblematic with respect to data dredging. Unfortunately, there is no definite answer to this question. Only a value of $H(c) = 0$ indicates completely stable ranking results that do not leave any room for data dredging (which is not realistic in practice). However, regarding the results in Figure 3.6 and 3.7, a value of $H(c) = 25$ for all $c \leqslant c_{max}$ (e.g. $c_{max} = 100$) might be acceptable. It has to be noted that this is only a rule of thumb that is not based on any objective criterion and might vary depending on the respective research question.



(a) $\rho = 0$

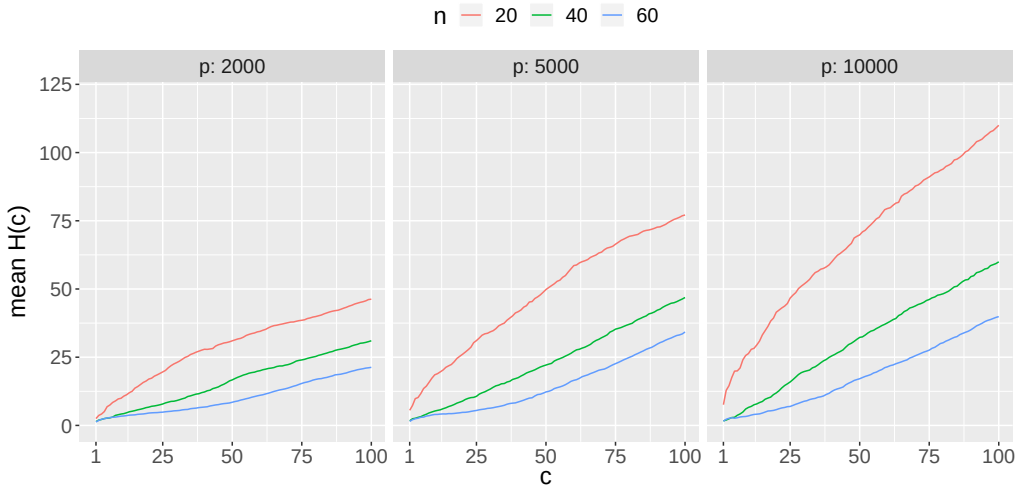


(b) $\rho = 0.4$

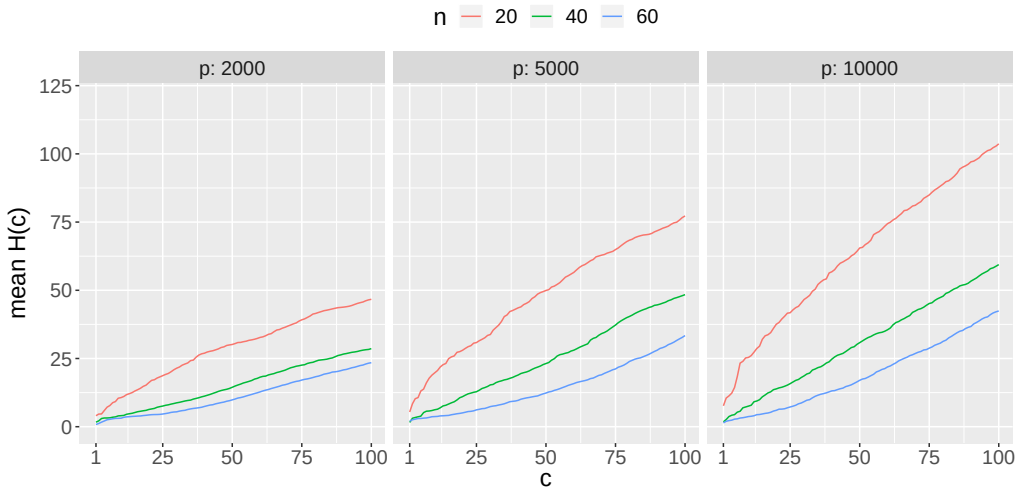


(c) $\rho = 0.8$

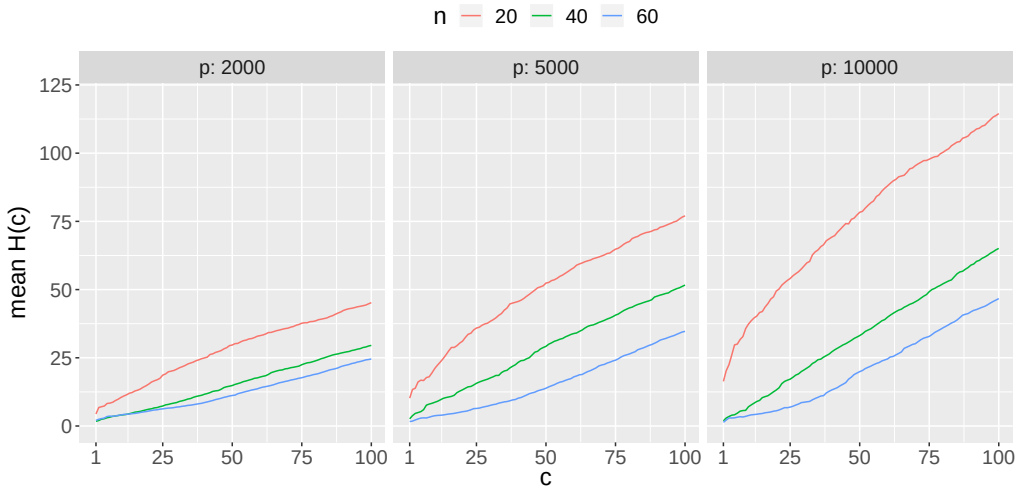
Figure 3.6.: Overall data dredging potential quantified by $H(c)$ for 27 simulated data sets with $n \in \{20, 40, 60\}$, $p \in \{2000, 5000, 10000\}$ and $\rho \in \{0, 0.4, 0.8\}$.



(a) $\rho = 0$



(b) $\rho = 0.4$



(c) $\rho = 0.8$

Figure 3.7.: Mean of $H(c)$ over 20 simulated data sets generated for each combination of $n \in \{20, 40, 60\}$, $p \in \{2000, 5000, 10000\}$ and $\rho \in \{0, 0.4, 0.8\}$.

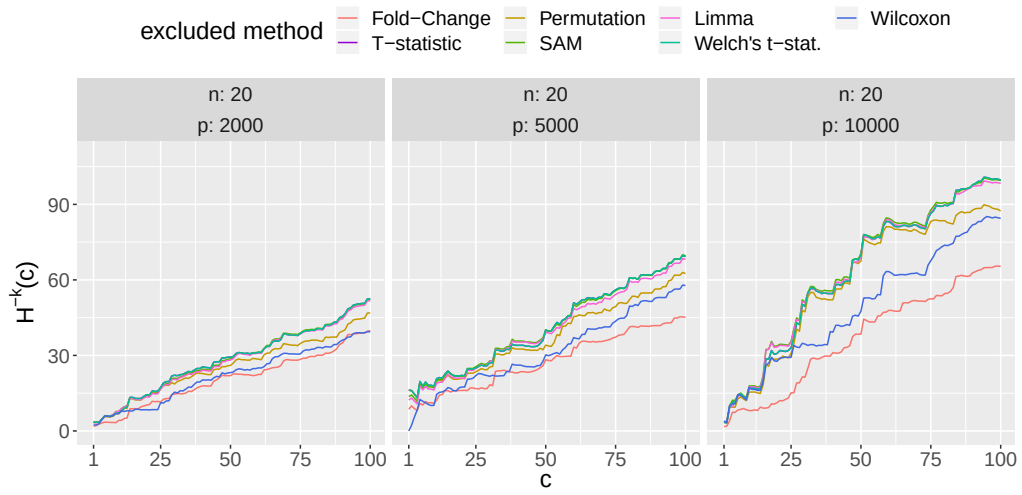


Figure 3.8.: Overall data dredging potential that arises when the variable rankings are performed without using method $k = 1, \ldots, 7$. The method that yields the smallest value of $H^{-k}(c)$ is the method that yields the highest increase in overall data dredging potential when added to the set of ranking methods. The figure shows three simulated data sets with $\rho = 0$, $n = 20$ and $p = \{2000, 5000, 10000\}$.

Step 3. Identify the method that yields the highest increase in data dredging potential

The last step aims to assess the increase in overall data dredging potential that arises if method $k = 1, \ldots, 7$ is used in addition to the other six methods. For this purpose, $H^{-k}(c)$ is calculated for each $k = 1, \ldots, 7$. As defined in Chapter 2, $H^{-k}(c)$ is the value of $H(c)$ that arises when the variable rankings are performed without considering method k.

Figure 3.8 displays $H^{-k}(c)$, $c \in [1, 100]$, for three simulated data sets with $n = 20$, $p \in \{2000, 5000, 10000\}$ and $\rho = 0$. It is striking that in each panel and for nearly all values of c, $H^{-FC}(c)$ yields the smallest value. This means that the fold-change criterion causes the highest increase in data dredging potential when it is used in addition to the other six methods. Two other ranking methods that noticeably increase the overall data dredging potential are the Wilcoxon statistic and the permutation test. The remaining four methods (Limma, SAM, t-statistic and Welch's t-statistic) yield almost the same values of $H^{-k}(c)$ for each value of c and p. Regarding the impact of the number of variables, the difference between $H^{-FC}(c)$ and the other values of $H^{-k}(c)$ is especially high for $p = 10000$. The same applies to $H^{-Wilcox}(c)$.

In principle, there are two possible reasons why a method increases the data dredging potential of a ranking result more than other methods: Either it assigns much higher ranks than the other methods or it assigns much smaller ranks. For the ranking results with $r_j^{best} \leqslant 100$ that are considered in Figure 3.8, it is likely that the latter case applies. This can be seen in Table 3.2, which was shown above. The data set in Table 3.2 corresponds to the left panel in Figure 3.8. It was stated that for all ranking results listed in the table,

r_j^{best} was exclusively assigned by the fold-change criterion, the Wilcoxon statistic or the permutation test. Hence, it is plausible that this is also true for the ranking results with $r_j^{best} \leqslant 100$ that are shown in Figure 3.8.

As stated in Chapter 2, the method that yields the highest increase in overall data dredging potential does not necessarily assign incorrect ranks. However, concerning the present case, this might be valid at least for the fold-change criterion. As stated in Section 3.2, this ranking method is very naive since it considers the differences in means only and does not take the variances into account. Thus, it can be expected that this methods yields incorrect ranking results (Boulesteix and Slawski, 2009).

The figures for the remaining data sets provide similar results and can be found in the appendix (Figures A.9-A.11).

4. Real data application with binary outcome

This chapter illustrates the framework introduced in Chapter 2 by applying it to a real data set with binary outcome. Sections 4.1 and 4.2 describe the data set and the ranking methods used for the application. Section 4.3 presents the results, which consist of variable rankings with respect to data dredging potential and the assessment of the overall data dredging potential.

4.1. Data

The data set $\mathcal{D}(\boldsymbol{x}, \boldsymbol{y})$ used throughout this chapter consists of a matrix $\boldsymbol{x} = (x_{ij})_{\substack{i=1,\ldots,114 \\ j=1,\ldots,12625}}$ and a vector $\boldsymbol{y} = (y_1, \ldots, y_{114})^\top$. The matrix $\boldsymbol{x}$ contains expression levels of 12625 genes from 114 individuals with acute lymphoblastic leukemia (ALL). For each individual, vector $\boldsymbol{y}$ stores the information if remission was achieved (99 individuals) or not (15 individuals). The data are available from the Bioconductor `ALL` package (Chiaretti et al., 2004; Li, 2018). Originally, 128 individuals were included in the data set but for 14 individuals, the remission status is not available.

4.2. Ranking methods

Since the data set corresponds to a two-group setting (remission vs. no remission), the variable rankings can be performed using the same seven methods as in the simulation (Chapter 3). To study the impact of the number of methods, the results are analysed for $m = 7$ and $m = 4$. In the latter case, the ranking methods based on the Wilcoxon statistic, the fold-change criterion and the permutation test are omitted. For the simulated data in Chapter 3, these three methods were shown to yield the highest increase in data dredging potential when added to the set of ranking methods. Note that omitting other ranking methods could possibly yield other results.
Table 4.1 gives an overview of the ranking methods used for $m = 7$ and $m = 4$. See Chapter 3 for a description of the methods.

4.3. Results

In this section, the ranking results of the ALL data set are analysed with respect to their data dredging potential. The procedure is the same as in Chapter 3: First, the genes

Table 4.1.: Overview of ranking methods used for the data application with binary outcome.

$m = 7$		$m = 4$	
k	method	k	method
1	Fold-change	1	T-statistic
2	T-statistic	2	SAM
3	Permutation test	3	Limma
4	SAM	4	Welch's t-statistc
5	Limma		
6	Welch's t-statistc		
7	Wilcoxon statistic		

are ranked with respect to their data dredging potential. Then, the overall data dredging potential is assessed.

Rankings with respect to data dredging potential

As suggested in Chapter 2, a combination of $h(\boldsymbol{r}_j; \alpha)$ and a cutoff c is used to rank the genes with respect to the data dredging potential of their ranking results. Parameter α varies with $\alpha \in \{0, 0.5, 0.8\}$ to illustrate its impact on the data dredging potential. The cutoff value is set to $c = 100$.

Table 4.2 presents the top-10 ranking results with the highest data dredging potential for $\alpha \in \{0, 0.5, 0.8\}$ and $m = 7$. First of all, it can be seen that compared to the simulation in Chapter 3, all ranking results in Table 4.2 yield very high values of $h(\boldsymbol{r}_j; \alpha)$. For almost all ranking results in Table 4.2, this is caused by the fold-change criterion, whose assigned ranks differ a lot from the ranks of the other methods. As already stated in Chapter 3, the fold change criterion is a very naive procedure that can be expected to yield incorrect ranking results.

The ranking result with the highest value of $h(\boldsymbol{r}_j; 0)$ is gene *38355_at*. For this gene, $h(\boldsymbol{r}_j; 0)$ takes a value of 4123.43, which means that the difference between the average and the best rank of gene *38355_at* is equal to 4123.43. For $\alpha = 0.5$ and $\alpha = 0.8$, this value is reduced to 601.46 and 189.49, respectively.

The only two values of $h(\boldsymbol{r}_j; \alpha)$ that remain constant for each α are the values of gene *36769_at* and *38124_at*. This is due to the fact that their best ranks are equal to 1. However, they are among the top-10 ranking results with the highest data dredging potential only if α is set to 0.8. In general, Table 4.2 reveals that for $\alpha > 0$, the ranking results with the highest data dredging potential tend to have smaller values of r_j^{best} than for $\alpha = 0$. As explained in Chapter 2, this is because for $\alpha > 0$, smaller values of r_j^{best} are considered as more relevant for data dredging, whereas for $\alpha = 0$, the actual value of r_j^{best} is not taken into account (except that $r_j^{best} \leqslant 100$).

The impact of α on the data dredging potential is additionally illustrated in Figure 4.1,

Table 4.2.: Top-10 lists of ranking results with the highest data dredging potential with $m = 7$, $c = 100$ and $\alpha \in \{0, 0.5, 0.8\}$. Empty cells denote that the ranking result is not among the top-10. For each ranking result, r_j^{best} is highlighted in red.

gene	$\boldsymbol{r}_j$	$h(\boldsymbol{r}_j;\alpha)$			rank (w.r.t. $h(\boldsymbol{r}_j;\alpha)$)		
		$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$
38355_at	(47 4724 5371 3032 4641 5271 6107)	4123.43	601.46	189.49	1	1	3
38319_at	(77 4229 3445 2668 4153 4982 7059)	3724.86	424.49	115.32	2	3	
37006_at	(60 3831 3552 2363 3737 4115 4303)	3077.29	397.28	116.32	3	5	
41214_at	(28 2990 3114 1722 2926 3052 1577)	2173.29	410.71	151.14	4	4	4
1914_at	(214 1990 1429 1175 1932 61 7858)	2033.14	260.32	75.84	5		
35192_at	(2080 2244 2003 1984 2214 100 2895)	1831.43	183.14	46.00	6		
41356_at	(85 1976 2226 1089 1919 3045 2909)	1807.71	196.07	51.71	7		
32649_at	(24 2275 2252 1225 2219 2325 2184)	1762.29	359.73	138.65	8	6	8
36536_at	(63 2454 2026 1390 2389 3036 1240)	1736.71	218.81	63.13	9		
35940_at	(45 779 1161 378 740 2217 6492)	1642.43	244.84	78.15	10		
38514_at	(8 1661 2075 787 1609 2376 1426)	1412.29	499.32	267.58		2	1
39318_at	(6 1036 1191 446 997 1029 1307)	852.86	348.18	203.40		7	2
39389_at	(17 1817 1816 914 1753 1444 2199)	1405.86	340.97	145.74		8	6
38242_at	(26 1816 1691 937 1751 2039 3414)	1641.71	321.97	121.15		9	10
40775_at	(27 1760 1685 909 1695 2273 1939)	1442.71	277.65	103.30		10	
36769_at	(978 4 4 20 4 20 1)	146.29	146.29	146.29			5
38124_at	(137 194 272 120 180 1 76)	139.00	139.00	139.00			7
34098_f_at	(160 422 667 220 409 3 183)	291.86	168.50	121.19			9

which consists of three panels that plot r_j^{best} against $h(\boldsymbol{r}_j; 0)$. In each panel, the top-10 ranking results shown in Table 4.2 are highlighted for $\alpha \in \{0, 0.5, 0.8\}$. Again, it can be observed that the average value of r_j^{best} for the top-10 results tends to decrease with increasing α.

The top-10 ranking results with the highest data dredging potential for $m = 4$ can be found in the appendix (Table A.1 and Figure A.12). Although the fold-change criterion is omitted for $m = 4$, the resulting values of $h(\boldsymbol{r}_j; \alpha)$ are still high compared to the simulation results.

Overall data dredging potential

The assessment of the overall data dredging potential is conducted in three steps as suggested in Chapter 2.

Step 1. Consider the distribution of ranks and the number of variables with $r_j^{best} \leqslant c$

To get a first impression of the overall data dredging potential, Figure 4.2 presents the number of variables with $r_j^{best} \leqslant c$ for $c \in [1, 100]$ and $m \in \{4, 7\}$. Ideally, the number of variables with $r_j^{best} \leqslant c$ should be equal to c for each $c \in [1, p]$. In this case, the overall data dredging potential would be non-existent because all methods assign the same rank to each variable. When looking at Figure 4.2, it is obvious that this is not the case for the ALL ranking results. For $m = 7$ and each $c \in [1, 100]$, the number of variables with

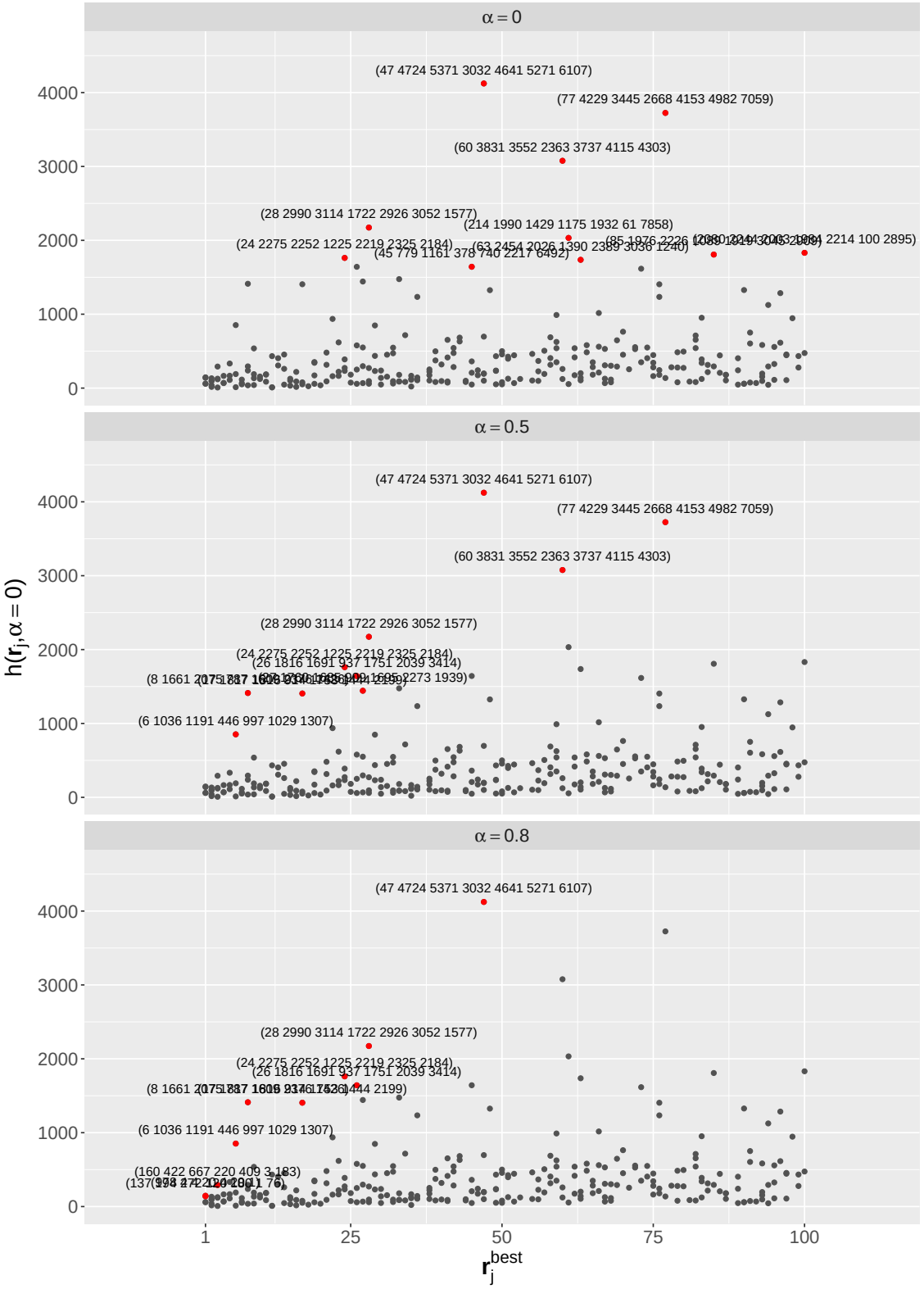


Figure 4.1.: Data dredging potential of ranking results with $r_j^{best} \leqslant 100$ for $m = 7$. In each panel, the top-10 ranking results with the highest data dredging potential for $\alpha \in \{0, 0.5, 0.8\}$ are highlighted in red.

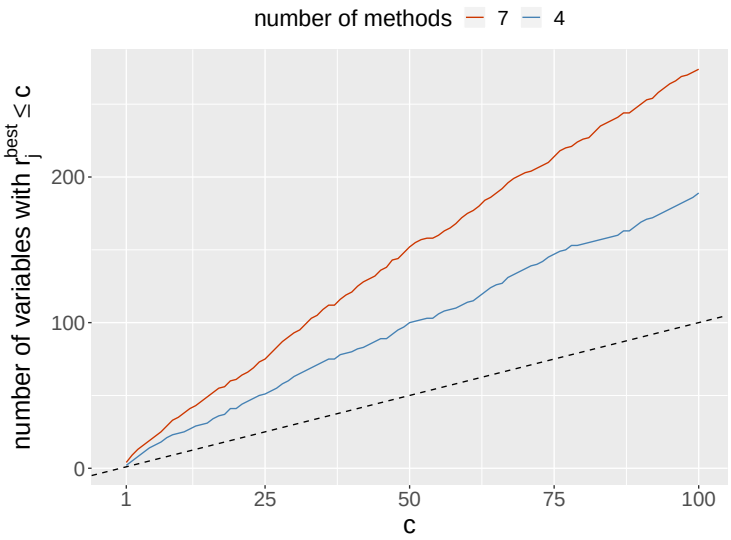


Figure 4.2.: Number of variables with $r_j^{best} \leqslant c$, $c \in [1, 100]$ and $m \in \{4, 7\}$. If all methods yield the same rank for each variable, the number of variables with $r_j^{best} \leqslant c$ is equal to c (dashed line).

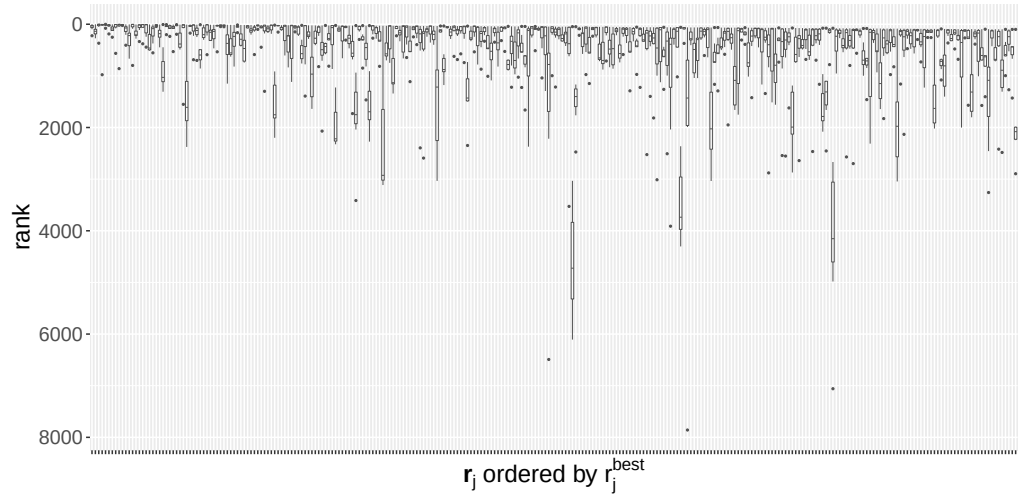
$r_j^{best} \leqslant c$ is approximately three times the value of c. Even for $m = 4$, which can be expected to yield more stable ranking results than $m = 7$, the number of variables with $r_j^{best} \leqslant c$ is almost twice the value of c.

As described in Chapter 2, another approach to get a first impression of the overall data dredging potential is to generate boxplots showing the distribution of each $\boldsymbol{r}_j$. Figure 4.3a displays the corresponding boxplots for all ranking results with $r_j^{best} \leqslant 100$ and $m = 7$. It reveals that the variability within the ranking results is in general very high. Many ranking results include ranks that are larger than 1000; a few ranking results even include ranks that are larger than 5000. For $m = 4$ (Figure 4.3b), the variability within the ranking results is considerably smaller. However, the range of many ranking results is still larger than 500, which indicates a high overall data dredging potential.

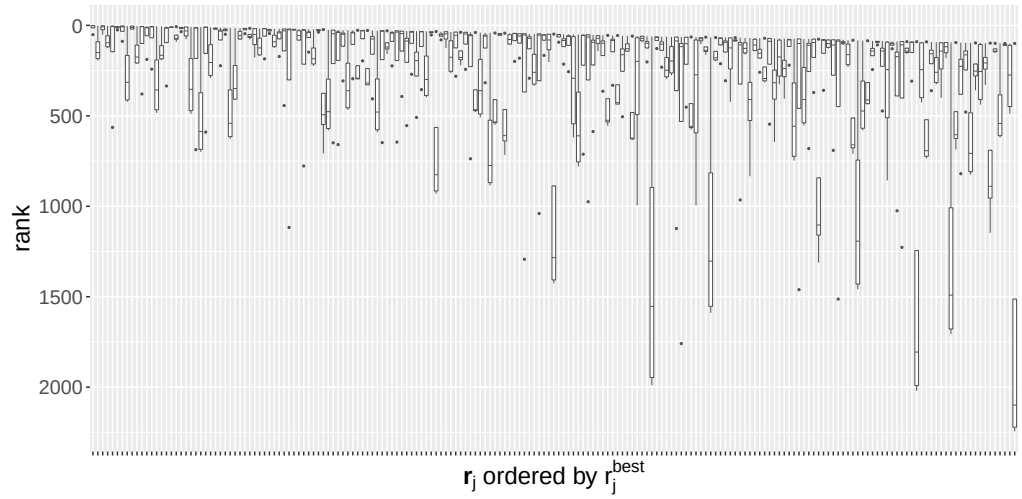
Step 2. Quantify the overall data dredging potential using $H(c)$

In the second step, the overall data dredging potential is quantified using $H(c)$, which is an aggregated version of $h(\boldsymbol{r}_j)$.

The resulting values of $H(c)$ for $c \in [1, 100]$ and $m \in \{4, 7\}$ are presented in Figure 4.4. Firstly, consider the values of $H(c)$ for $m = 7$. For $c = 1$, $H(c)$ is approximately equal to 100. Accordingly, the mean difference between the average and the best rank of variables with $r_j^{best} = 1$ is approximately equal to 100. This value increases rapidly with c, which indicates that the variability within the ranking results increases with r_j^{best}. For $m = 4$, $H(c)$ is considerably smaller than for $m = 7$, especially with increasing values of c. Nevertheless, both $m = 4$ and $m = 7$ yield higher values of $H(c)$ than the simulated



(a) $m = 7$



(b) $m = 4$

Figure 4.3.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ for $m \in \{4, 7\}$.

ranking results in Chapter 3. Specifically, taking the proposed threshold of $H(c) = 25$ as a basis, the overall data dredging potential of the ranking results can be considered as very problematic, even for $m = 4$.

Overall, Figure 4.4 confirms the findings from Step 1: For both $m = 4$ and $m = 7$, the ranking results from the ALL data set show a considerably higher overall data dredging potential than the simulated ranking results in Chapter 3 (which were generated by the same seven ranking methods). Moreover, the ranking results from the ALL data set have a smaller but non-negligible overall data dredging potential if the Wilcoxon statistic, the fold-change criterion and the permutation test are omitted.

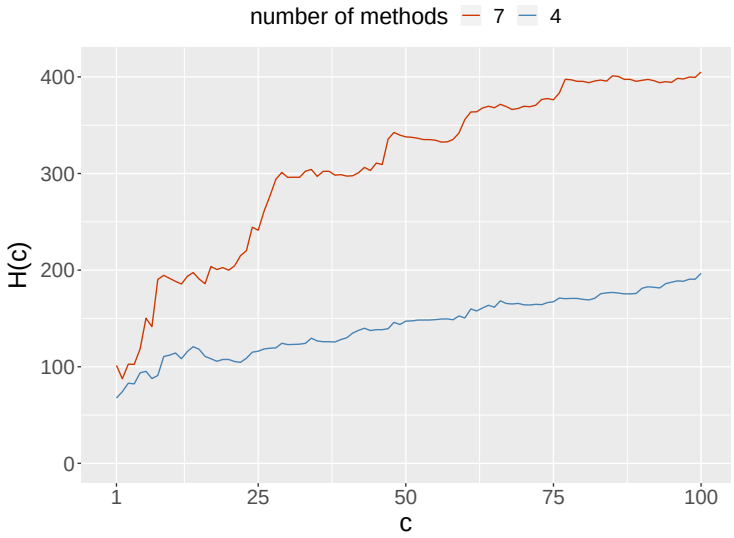


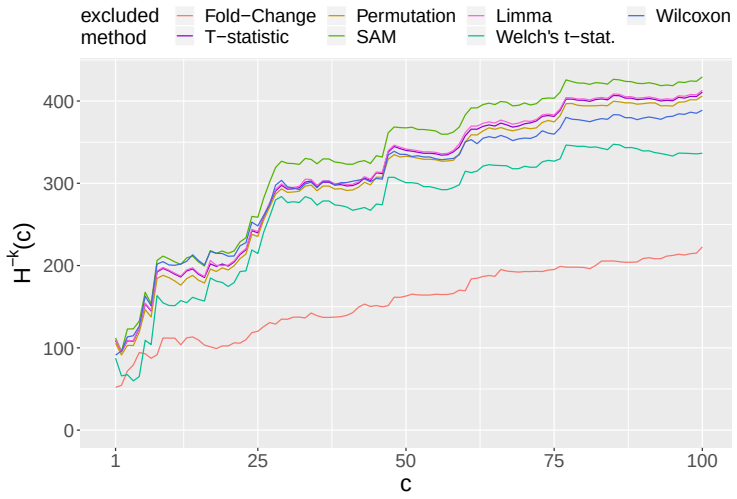
Figure 4.4.: Overall data dredging potential quantified by $H(c)$ for $m \in \{4, 7\}$.

Step 3. Identify the method that yields the highest increase in data dredging potential

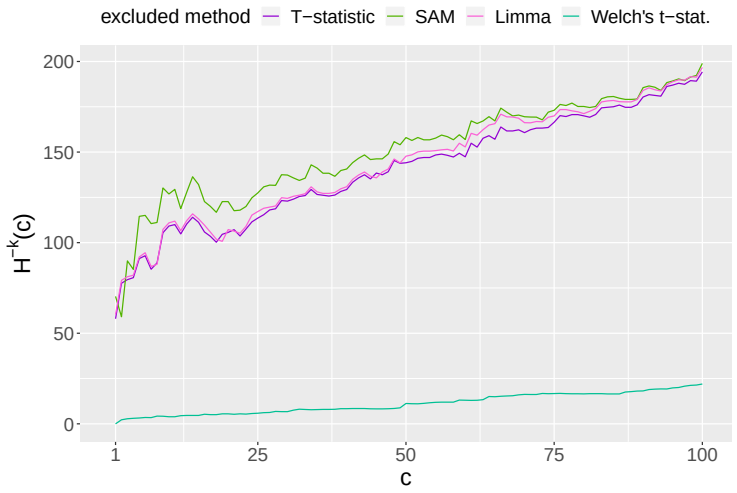
In the last step, the ranking methods are analysed with respect to their contribution to the overall data dredging potential. For this purpose, $H^{-k}(c)$ is calculated for each $k = 1, \ldots, m$ with $m \in \{4, 7\}$. As explained in Chapter 2, $H^{-k}(c)$ is the value of $H(c)$ that arises if method k is omitted from the ranking analyses. The method that yields the smallest value of $H^{-k}(c)$ is the method that yields the highest increase in overall data dredging potential if it is added to the set of ranking methods.

Figure 4.5a presents the results for $m = 7$ and $c \in [1, 100]$. In Table 4.2, it was shown that the high values of $h(\boldsymbol{r}_j; \alpha)$ are mainly caused by the fold-change criterion, whose assigned ranks differ a lot from the other methods. Figure 4.5a reveals that this is not just applicable for the top-10 ranking results with the highest data dredging potential but also for other ranking results with $r_j^{best} \leqslant 100$: The method with the smallest value of $H^{-k}(c)$ for almost all values of c is the fold-change criterion. Consequently, the fold-change criterion yields the highest increase in overall data dredging potential when it is added to the set of ranking methods. Moreover, it can be observed that $H^{-FC}(c)$ in Figure 4.5a is approximately equal to $H(c)$ for $m = 4$ in Figure 4.4. This indicates that the difference in overall data dredging potential between $m = 4$ and $m = 7$ that was revealed in Step 1 and 2 is mainly due to the fact that the fold-change criterion is omitted for $m = 4$. As already stated in Chapter 3, the fold change criterion is a very naive procedure that should be used with caution. Compared to the fold-change criterion, the Wilcoxon statistic and the permutation test do not seem to reduce the overall data dredging potential considerably when they are omitted.

Figure 4.5b displays the values of $H^{-k}(c)$ for $m = 4$. It is immediately apparent that Welch's t-statistic yields the smallest value of $H^{-k}(c)$ for $c \in [1, 100]$. In contrast, the



(a) $m = 7$



(b) $m = 4$

Figure 4.5.: Overall data dredging potential that arises when the variable rankings are performed without using method $k = 1, \ldots, m$, $m \in \{4, 7\}$. The method with the smallest value of $H^{-k}(c)$ is the method that yields the highest increase in overall data dredging potential when added to the set of ranking methods.

values of $H^{-k}(c)$ for the remaining three methods are approximately equal to the values of $H(c)$ for $m = 4$ in Figure 4.4. This result is probably due to the fact that SAM and Limma are modified versions of the t-statistic. Thus, all three methods are based on a test that assumes that the variances in both groups are equal. In contrast, Welch's t-statistic assumes unequal variances in both groups (Jeanmougin et al., 2010). Consequently, it is plausible that the rankings generated by t-statistic, SAM and Limma are very similar, which in turn leads to a low overall data dredging potential if Welch's t-statistic is omitted. However, that this does not imply that Welch's t-statistic yields wrong ranking results.

5. Real data application with survival outcome

In Chapter 3 and 4, the framework for the quantification of data dredging potential in ranking analyses was applied to data sets with binary outcome. The present chapter analyses the data dredging potential of ranking results in the survival context. Section 5.1 introduces the data set that is used to illustrate the framework. Section 5.2 then describes the considered ranking methods. In Section 5.3, the ranking results are analysed with respect to their data dredging potential.

5.1. Data

For the ranking analyses with survival outcomes, a data set of the form $\mathcal{D}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\delta})$ is considered, where $\boldsymbol{x} = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ store n independent observations of $X_1, \ldots, X_p$, Y and δ. Variable Y denotes the follow-up time (or censored survival time), which is defined as the minimum of the underlying survival time T and the censoring time C, i.e. $Y = \min(T, C)$. Variable δ denotes the censoring indicator and is defined as $\delta = \mathbb{1}(T \leqslant C)$, where $\mathbb{1}(\cdot)$ is the indicator function. As in the previous chapters, $X_1, \ldots, X_p$ denote the variables that are ranked with respect to their importance for the outcome. (Edelmann et al., 2019b; Chen et al., 2018)
The survival data set used throughout this chapter is based on the mantle cell lymphoma (MCL) study of the Lymphoma/Leukemia Molecular Profiling Project and is available at `http://llmpp.nih.gov/MCL` (Rosenwald et al., 2003). It contains the expression values of 8810 genes from 92 untreated MCL patients with no history of previous lymphoma. As in earlier applications (e.g. Edelmann et al., 2019b), the analysis will be restricted to genes that do not contain missing observations, which reduces the number of genes to 2480. In addition to the gene expression values, the data set contains the follow-up time and the censoring status of each patient. During the follow-up, 64 patients died of MCL; the other 28 patients were censored. The median follow-up time is 2.76 years.

5.2. Ranking methods

Since the outcome of interest is a survival time, the genes in the MCL data set cannot be ranked by the same methods as in Chapter 3 and 4. Instead, they are ranked by eight methods that are suitable for survival data. As in the previous applications, the considered methods differ with respect to their ranking criterion. Moreover, all methods

Table 5.1.: Overview of ranking methods used for the data application with survival outcome.

$m = 8$		$m = 4$	
k	method	k	method
1	COX	1	COX
2	SIS	2	CINDEX
3	CINDEX	3	CRCDCS
4	IPOD	4	RESI
5	RCDCS		
6	CRCDCS		
7	RESI		
8	BCORSIS		

are univariate. Accordingly, each method is based on a ranking criterion that is calculated for each variable X_j without considering the other variables $X_1, \ldots, X_{j-1}, X_{j+1} \ldots, X_p$. The variables are then ranked according to the value of the ranking criterion, where a high value indicates high importance for the response. In the context of univariate ranking methods for survival data, the ranking criterion is also referred to as *marginal utility*. (Edelmann et al., 2019b)

An overview of the considered ranking methods can be found in Table 5.1. All methods are implemented in the R package `MVS` (Edelmann et al., 2019a). As in the real data application with binary outcome (Chapter 3), the impact of the number of methods is studied by additionally analysing the ranking results for $m = 4$. For this purpose, four methods were randomly chosen from the set of ranking methods. The corresponding methods are also listed in Table 5.1.

The remainder of this section briefly reviews the methods that were used to generate the variable rankings:

- **(SIS-)COX**: The marginal utility of variable X_j is defined as the maximum of the partial likelihood of a Cox model with variable X_j as predictor (Fan et al., 2010). SIS is the abbreviation for *sure independence screening*, which is an approach by Fan and Lv (2008) that uses Pearson correlation for feature screening and was theoretically justified for linear models. COX is an adhoc solution to apply SIS to the Cox proportional-hazards model (Edelmann et al., 2019b).

- **SIS**: The marginal utility of variable X_j is determined by calculating the absolute value of the Pearson correlation between follow-up time Y and variable X_j. It was proposed by Saldana and Feng (2018) as a computational shortcut for the COX method. However, by considering the follow-up time Y instead of the survival time T, it does not correct appropriately for censoring (Edelmann et al., 2019b).

- **CINDEX**: This method uses Harrel's C-Index (C for concordance) (Harrell et al., 1996), which is a rank correlation statistic and thus not sensitive to outliers in the

data. The C-Index of variable X_j is given by

$$C(\boldsymbol{x}_{\cdot j}) = \frac{\sum_{i=1}^{n}\sum_{i'=1}^{n} \mathbb{1}(y_i > y_{i'})\mathbb{1}(x_{i'j} > x_{ij})\delta_{i'}}{\sum_{i=1}^{n}\sum_{i'=1}^{n} \mathbb{1}(y_i > y_{i'})\delta_{i'}}, \tag{5.1}$$

where $\boldsymbol{x}_{\cdot j} = (x_{1j}, \ldots, x_{nj})$ contains n observations of variable X_j. The marginal utility based on the C-Index is defined as:

$$\max(C(\boldsymbol{x}_{\cdot j}), C(-\boldsymbol{x}_{\cdot j})) \tag{5.2}$$

(Edelmann et al., 2019b).

- **IPOD**: In contrast to the methods above (COX, SIS and CINDEX), this method allows to detect non-monotone associations between the variables and the survival time. It is based on the consideration that the survival time T is independent of variable X_j if and only if the conditional survival function $S(t\,|X_j = x)$ equals the unconditional survival function $S(t)$ for almost all x (Edelmann et al., 2019b). Specifically, Hong et al. (2018) propose the *integrated power density* to assess the marginal utility. It is defined as:

$$\text{IPOD}_\gamma(t, r) = \int_0^t f^\gamma(s\,|X_j = r)ds, \tag{5.3}$$

where $r \in \{1, \ldots, R_j\}$ is a category of the (discretized) variable X_j, $\gamma > 0$ is a tuning parameter and $f(t|X_j = r)$ is the conditional density function of t, which is estimated in practice. For $\gamma = 1$, equation (5.3) is the conditional cumulative density function $F(t|X_j = x)$. The marginal utility of variable X_j is the maximum absolute difference of IPOD over all pairs $r_1, r_2 \in \{1, \ldots, R_j\}$:

$$\max_{r_1, r_2 \in \{r_1, \ldots, r_m\}} \sup_{t \in [0,\tau)} |\text{IPOD}_\gamma(t, r_i) - \text{IPOD}_\gamma(t, r_j)|, \tag{5.4}$$

where $\tau > 0$ is a fixed time point (e.g., the study duration). In the implementation of the R package `MVS`, the tuning parameter γ is set to 1.

The remaining four methods are based on the distance correlation coefficient or related measures. Thus, the concept of distance correlation is explained briefly based on Edelmann et al. (2019b).
Distance correlation is a powerful measure of dependence. It allows to detect arbitrary (including non-monotone) associations between two variables of arbitrary dimension and is 0 if and only if the variables are independent. The distance correlation coefficient between

two random variables U and V is defined as

$$\mathrm{dCor}(U,V) = \frac{\mathrm{dCov}(U,V)}{\sqrt{\mathrm{dCov}(U,U)\mathrm{dCov}(V,V)}}, \tag{5.5}$$

where $\mathrm{dCov}(\cdot,\cdot)$ denotes the *distance covariance*. It is defined as the non-negative square root of

$$\begin{aligned}\mathrm{dCov}^2(U,V) =& \mathrm{E}(|U-U'||V-V'|) + \mathrm{E}(|U-U'|)\mathrm{E}(|V-V'|) \\ &- 2\mathrm{E}(|U-U'||V-V''|),\end{aligned} \tag{5.6}$$

where U and V are again random variables and (U',V') and V'' are independent copies of (U,V) and V, respectively. For random vectors with $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^q$, the absolute value $|\cdot|$ in (5.6) is replaced by the corresponding Euclidean norm.
In practice, $\mathrm{dCor}(U,V)$ is estimated from a sample of (U,V). A consistent and unbiased estimator for $\mathrm{dCor}(U,V)$ can be found in Edelmann et al. (2019b).
The following methods use distance correlation:

- **RCDCS**: Chen et al. (2018) have proposed the *robust censored distance correlation screening*, which uses the squared distance correlation between the cumulative distribution functions of X_j and T as marginal utility:

  $$\mathrm{dCor}^2(F_j(X_j), F(T)). \tag{5.7}$$

  To estimate (5.7), $F_j(X_j)$ is replaced by the corresponding empirical distribution function and dCor by a suitable sample version. Additionally, $F(\cdot)$ is substituted by its Kaplan-Meier estimate and the survival time T (which cannot be observed in practice) by the follow-up time Y (Edelmann et al., 2019b; Chen et al., 2018). Edelmann et al. (2019b) note that by using Y instead of T, RCDCS does not appropriately correct for censoring.

- **CRCDCS**: In the same paper, Chen et al. (2018) have introduced a second method which they refer to as *composite robust censored distance correlation screening*. The idea behind this method is that if variable X_j has influence on the survival time T, there must exist some τ such that the τ-th quantile of T $(=Q_\tau(T))$ depends on X_j. This is equivalent to testing whether $W_\tau = \tau - \mathbb{1}(T \leqslant Q_\tau(T))$ and X_j are independent. The marginal utility of variable X_j measures the association between X_j and multiple quantiles $Q_{\tau_1}(T), \ldots, Q_{\tau_H}(T)$ simultaneously. It is defined as

  $$\mathrm{dCor}^2(F_j(X_j), (\tilde{W}_{\tau_1}, \ldots, \tilde{W}_{\tau_H})^\top)), \tag{5.8}$$

where $F_j(X_j)$ is the cumulative density function of X_j and $(\tilde{W}_{\tau_1}, \ldots, \tilde{W}_{\tau_H})^\top$ are weight-adjusted versions of $(W_{\tau_1} \ldots, W_{\tau_H})^\top$ that take censoring into account. See Chen et al. (2018) for a sample version of (5.8).

- **RESI**: This method has been proposed by Edelmann et al. (2019b). It circumvents the problem that the survival time T cannot be observed directly by using the distance correlation between X_j and the martingale residuals of a null Cox model as marginal utility. The martingale residual of the i-th individual is $\widehat{M}_i = \delta_i - \hat{\Delta}(Y_i)$, where $\hat{\Delta}(\cdot)$ is the Breslow estimate of the cumulative Baseline-hazard.

- **BCORSIS**: The *ball correlation sure independence screening* proposed by Pan et al. (2018) is similar to RCDCS but instead of distance correlation, it uses the closely related *Ball correlation* between T and X_j as marginal utility (Edelmann et al., 2019b). More information about the Ball correlation can be found in Pan et al. (2018).

For details on the methods, the reader is referred to the cited literature as well as to Edelmann et al. (2019b), which give a comprehensive overview of ranking methods in the survival context.

5.3. Results

In this section, the ranking results of the MCL data set are analysed with respect to their data dredging potential. As in the previous applications (Chapter 3 and 4), this includes rankings with respect to data dredging potential and the assessment of the overall data dredging potential.

Rankings with respect to data dredging potential

The genes are ranked with respect to their data dredging potential by using a combination of $c = 100$ and $h(\boldsymbol{r}_j; \alpha)$ with $\alpha \in \{0, 0.5, 0.8\}$. Table 5.2 shows the top-10 list of ranking results with the highest data dredging potential for each α and $m = 8$. First of all, it can be observed that the values of $h(\boldsymbol{r}_j; \alpha)$ are smaller than the values of $h(\boldsymbol{r}_j; \alpha)$ for the top-10 ranking results of the ALL data set (Table 4.2 in Chapter 4). However, as stated in Chapter 2, the maximum value of $h(\boldsymbol{r}_j; \alpha)$ depends on the number of variables and methods. In the present case, with $p = 2480$ and $m = 8$, the highest possible value of $h(\boldsymbol{r}_j; \alpha)$ is equal to $\frac{m-1}{m}(p-1) = 2169.125$ for each α, whereas the highest possible value of the ALL data set is 10820.57 ($m = 7$ and $p = 12625$).
As can be seen in Table 5.2, the variable with the highest data dredging potential for $\alpha = 0$ and $\alpha = 0.5$ is gene *16370*. The difference between its average rank and best rank is 1201.38 ($h(\boldsymbol{r}_j; 0) = 1201.38$). For $\alpha = 0.8$, the value of $h(\boldsymbol{r}_j; \alpha)$ is reduced to 31.18 since the best rank of gene *16370* is equal to 96.

Table 5.2.: Top-10 lists of ranking results with the highest data dredging potential with $m = 8$, $c = 100$ and $\alpha \in \{0, 0.5, 0.8\}$. Empty cells denote that the ranking result is not among the top-10. For each ranking result, r_j^{best} is highlighted in red.

gene	$\boldsymbol{r}_j$	$h(\boldsymbol{r}_j;\alpha)$			rank (w.r.t. $h(\boldsymbol{r}_j;\alpha)$)		
		$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$
16370	(1304 1875 1615 96 1819 937 729 2004)	1201.38	122.61	31.18	1	1	3
29821	(1131 1198 1523 94 1517 1623 1300 1444)	1134.75	117.04	29.95	2	3	5
31503	(407 78 915 1527 414 704 640 430)	561.38	63.56	17.20	3	7	
33558	(564 779 627 75 904 612 357 756)	509.25	58.80	16.10	4	8	
30828	(389 100 549 873 346 863 735 404)	432.38	43.24	10.86	5		
28805	(852 1183 555 30 292 307 269 172)	427.50	78.05	28.13	6	6	6
27810	(297 19 591 575 130 726 886 179)	406.38	93.23	38.54	7	4	2
28595	(321 52 582 930 317 590 355 333)	383.00	53.11	16.23	8	9	
26207	(305 23 580 970 250 588 396 134)	382.75	79.81	31.16	9	5	4
32501	(749 939 429 93 449 312 353 388)	371.00	38.47	9.88	10		
31037	(99 1 143 301 107 134 115 62)	119.25	119.25	119.25		2	1
17179	(369 846 487 44 268 238 250 207)	294.62	44.42	14.27		10	
27116	(127 20 60 81 3 79 140 16)	62.75	36.23	26.06			7
15936	(21 2 37 167 16 49 41 2)	39.88	28.20	22.90			8
16020	(87 8 155 113 72 199 194 87)	106.38	37.61	20.15			9
15886	(78 116 72 131 45 83 79 5)	71.12	31.81	19.63			10

As in the previous applications, the values of r_j^{best} in the top-10 list are on average higher for $\alpha = 0$ than for $\alpha > 0$. As an example, consider the top-10 list for $\alpha = 0.8$: Half of the ranking results in the top-10 list provide a r_j^{best} smaller than 9. This tendency regarding the impact of α is additionally illustrated in Figure 5.1, which displays all ranking results with $r_j^{best} \leqslant 100$ and highlights the top-10 ranking results for each $\alpha \in \{0, 0.5, 0.8\}$.
When looking at the best ranks in Table 5.2, it is striking that these are almost exclusively assigned by SIS and IPOD. In particular, note that the two genes that are identified as most relevant by SIS (gene *31037* and *15936*) are among the top-10 ranking results with the highest data dredging potential for $\alpha = 0.8$.
The top-10 ranking results with the highest data dredging potential for $m = 4$ yield considerably smaller values of $h(\boldsymbol{r}_j;\alpha)$. Corresponding results can be found in the appendix (Table A.2 and Figure A.13).

Overall data dredging potential

This section aims to provide an overview about how unstable the ranking results of the MCL data set are with respect to data dredging. As in the previous applications, the overall data dredging potential is assessed in three steps.

Step 1. Consider the distribution of ranks and the number of variables with $r_j^{best} \leqslant c$

In the first step, the overall data dredging potential is assessed graphically by considering the number of variables with $r_j^{best} \leqslant c$ and the distribution of each ranking result.

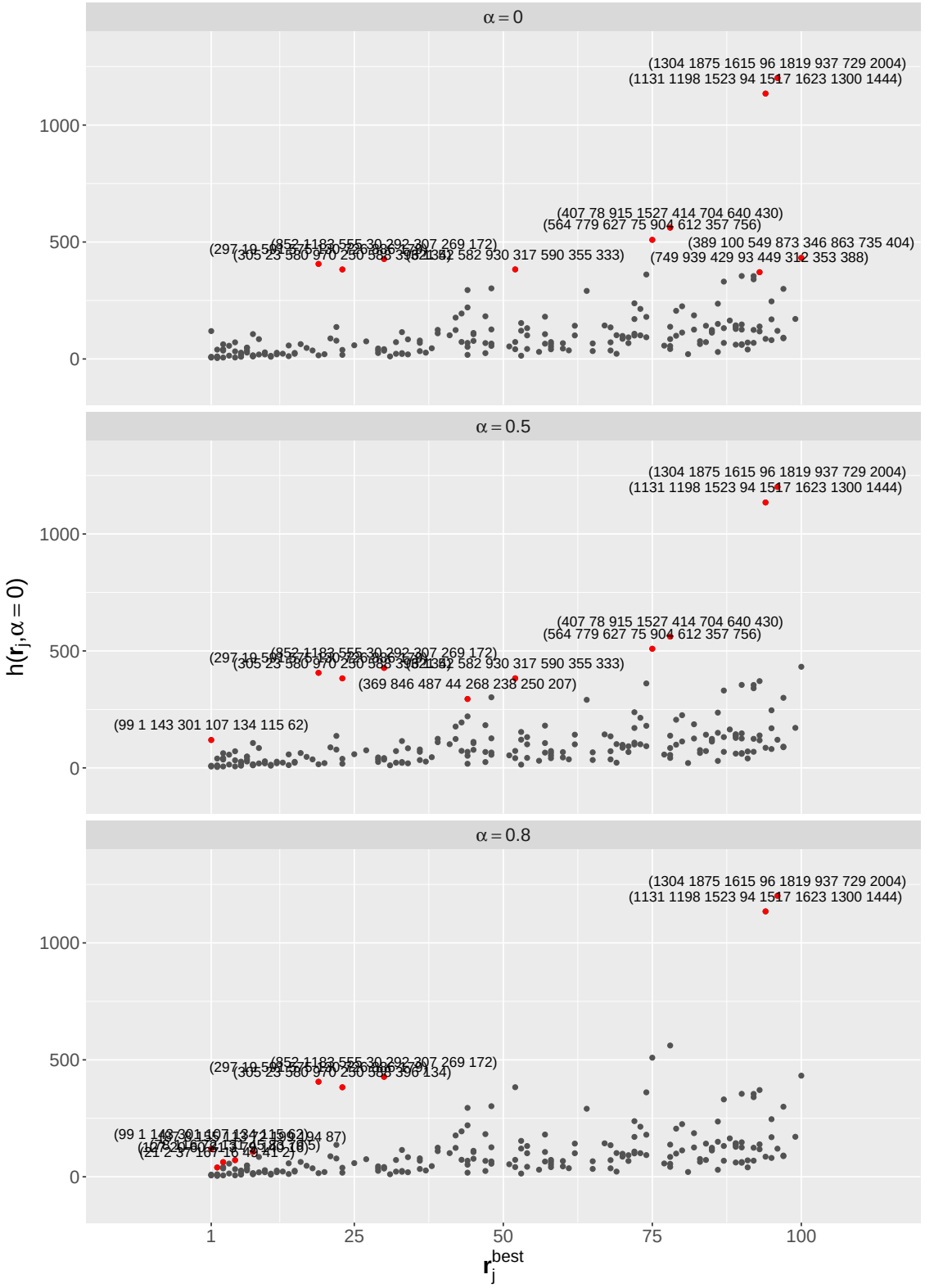


Figure 5.1.: Data dredging potential of ranking results with $r_j^{best} \leqslant 100$ for $m = 8$. In each panel, the top-10 ranking results with the highest data dredging potential for $\alpha \in \{0, 0.5, 0.8\}$ are highlighted in red.

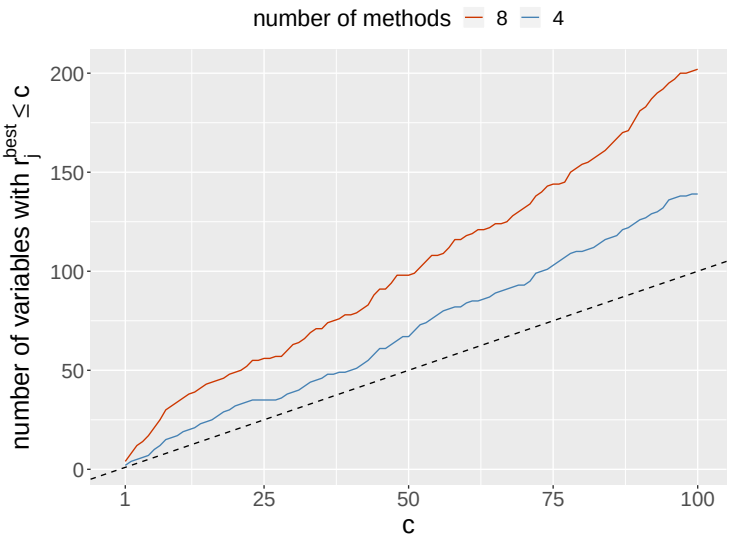
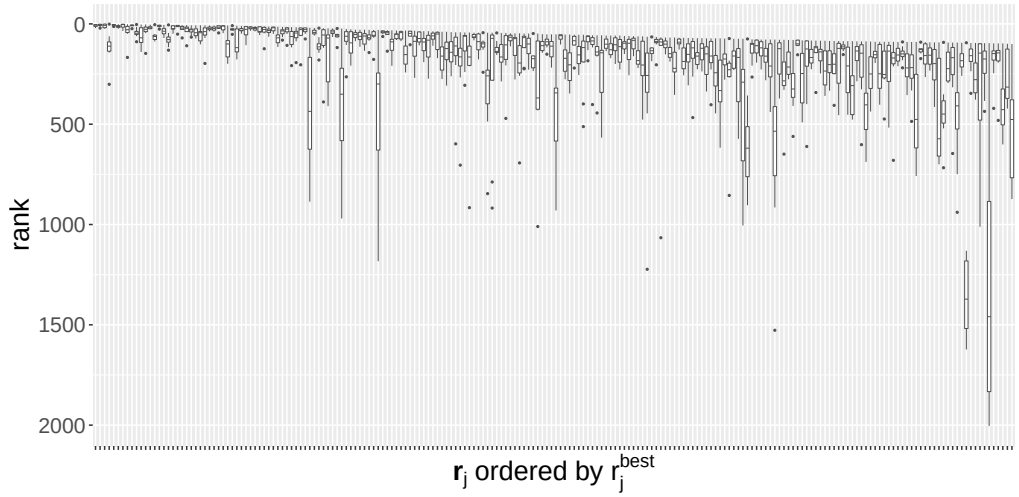


Figure 5.2.: Number of variables with $r_j^{best} \leqslant c$, $c \in [1, 100]$ and $m \in \{4, 8\}$. If all methods yield the same rank for each variable, the number of variables with $r_j^{best} \leqslant c$ is equal to c (dashed line).

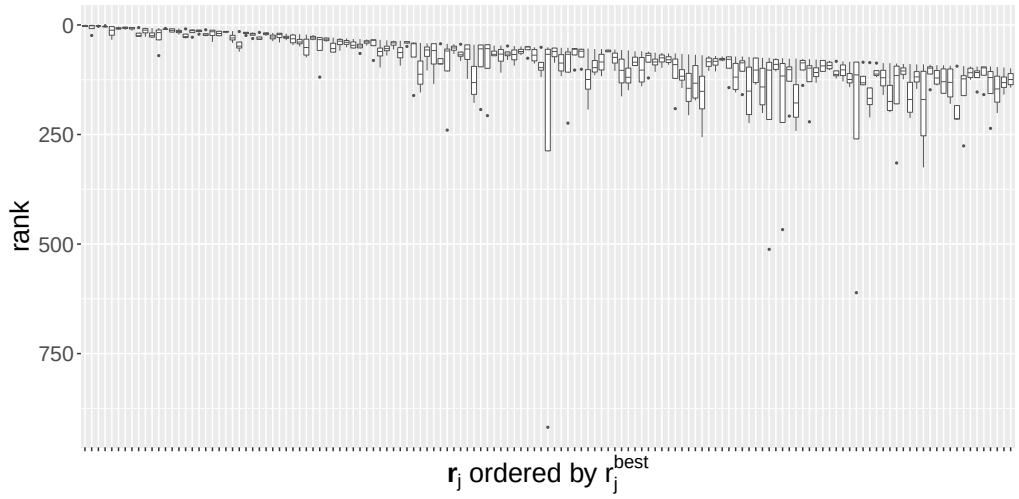
As can be seen from Figure 5.2, for $m = 8$, the number of variables with $r_j^{best} \leqslant c$ is approximately twice the value of c. For $m = 4$, this number is considerably smaller.
Figure 5.3a illustrates the distribution of each ranking result with $r_j^{best} \leqslant 100$ for $m = 8$. It reveals that many ranking results include ranks that are larger than 500. One ranking result even include a rank that is larger than 2000. In fact, this is the ranking result of gene *16370*, which was identified as the ranking result with the highest data dredging potential for $\alpha \in \{0, 0.5\}$ in Table 5.2. Furthermore, Figure 5.3a shows that the variability within the ranking result tends to increase with r_j^{best}.
The tendency regarding the relation between the value of r_j^{best} and the variability within the ranking results can also be observed in Figure 5.3b, which shows the distribution of each $\boldsymbol{r}_j$ for $m = 4$. However, the variability within the ranking results for $m = 4$ is smaller than for $m = 8$; only a few ranking results include ranks that are larger than 250.

Step 2. Quantify the overall data dredging potential using $H(c)$

In the second step, the overall data dredging potential is quantified by $H(c)$. As in the previous applications, $H(c)$ is considered for $c \in [1, 100]$. The results for $m \in \{4, 8\}$ are presented in Figure 5.4. In principle, they confirm the findings from Step 1: The variability within the ranking results increases with c and is considerably smaller for $m = 4$ than for $m = 8$. However, in contrast to the figures in Step 1, $H(c)$ additionally allows to quantify the mean variability in the ranking results with respect to r_j^{best}. In Chapter 3, a value of $H(c) = 25$ was proposed as a threshold for differentiating between problematic and unproblematic overall data dredging potential. Taking this rule of thumb as a basis, the



(a) $m = 8$



(b) $m = 4$

Figure 5.3.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ for $m \in \{4, 8\}$.

ranking results for $m = 8$ show an overall data dredging potential that can be considered as problematic. For $m = 4$ and $c \leqslant 75$, the overall data dredging potential does not seem to be particularly problematic.

Step 3. Identify the method that yields the highest increase in data dredging potential

In the final step, the ranking methods are analysed with respect to their contribution to the overall data dredging potential. As in the previous applications, this is done by calculating $H^{-k}(c)$ for each $k = 1, \ldots, m$, where $H^{-k}(c)$ is the value of $H(c)$ that arises if method k is not considered for the ranking analyses.
The results for $m = 8$ and $c \in [1, 100]$ are displayed in Figure 5.5a. It shows that SIS yields the smallest value of $H^{-k}(c)$ for almost all values of c. Consequently, the overall

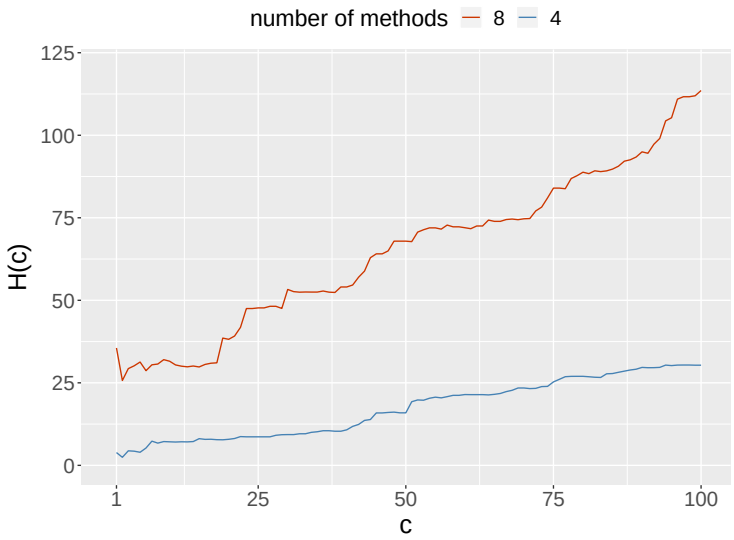
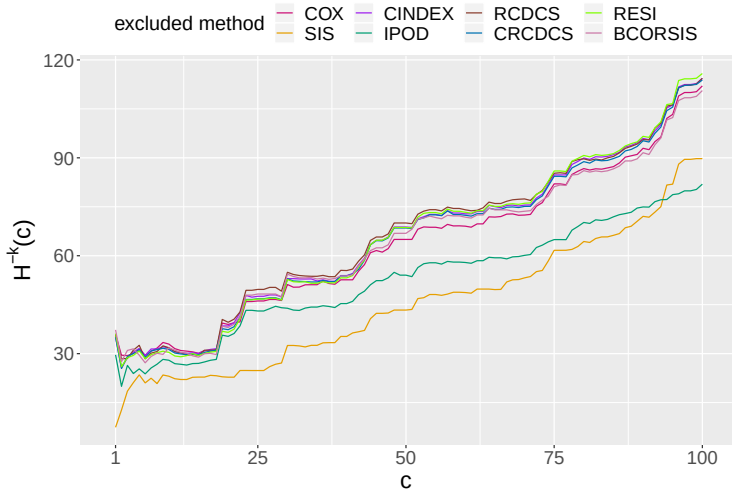


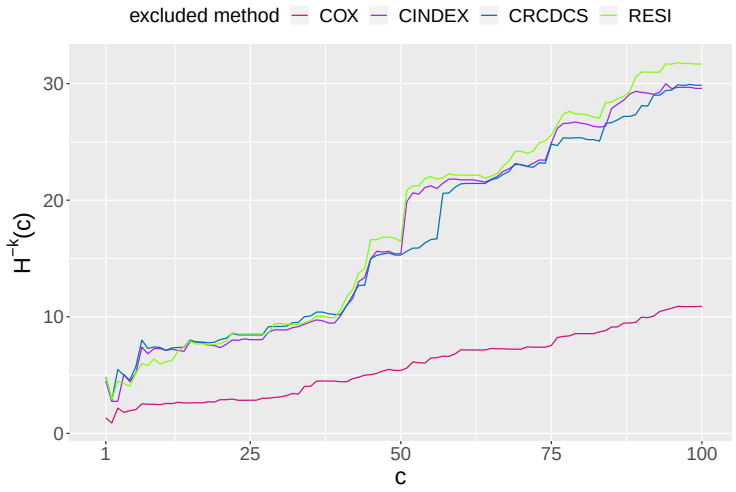
Figure 5.4.: Overall data dredging potential quantified by $H(c)$ for $m \in \{4, 8\}$.

data dredging potential increases the most if SIS is added to the set of ranking methods. The reason for this may be that SIS does not properly correct for censoring. However, this also applies to RCDCS, which does not appear to increase the overall data dredging potential a lot if added to the set of ranking methods. Apart from SIS, the only method that increases the overall data dredging potential noticeably is IPOD. Note that both SIS and IPOD were already shown to yield much smaller ranks than the other methods with regard to the top-10 ranking results with the highest data dredging potential (Table 5.2). Although it is uncertain why SIS and IPOD differ from the other ranking methods, they should be used with caution. However, as stated in Chapter 2, this does not necessarily imply that SIS and COX yield wrong ranking results.

Figure 5.5b presents the results for $m = 4$. It reveals that the overall data dredging potential decreases the most if COX is omitted from the set of ranking methods. This indicates that CINDEX, CRCDCS and RESI yield very similar ranking results. However, as shown in Figure 5.4, even if COX is added to the set of ranking methods (i.e. for $m = 4$), the overall data dredging potential is very low, which indicates that all four methods yield similar ranking results.



(a) $m = 8$



(b) $m = 4$

Figure 5.5.: Overall data dredging potential that arises when the variable rankings are performed without using method $k = 1, \ldots, m$, $m \in \{4, 8\}$. The method with the smallest value of $H^{-k}(c)$ is the method that yields the highest increase in overall data dredging potential when added to the set of ranking methods.

6. Conclusion

The aim of this thesis was to provide a framework that allows to quantify the data dredging potential in ranking results. As a first step, a measure $h(\boldsymbol{r}_j)$ was introduced that quantifies the variability within the ranking result of variable X_j with respect to its best rank r_j^{best}. This measure can be interpreted as the difference between the average rank and the best rank of variable X_j. A drawback of $h(\boldsymbol{r}_j)$ is that it does not take the actual value of r_j^{best} into account. Thus, two modifications of $h(\boldsymbol{r}_j)$ were proposed. The first one is $h(\boldsymbol{r}_j; \alpha)$, which weights $h(\boldsymbol{r}_j)$ according to r_j^{best} and a parameter α. The second approach consists of choosing a cutoff value c that assigns a data dredging potential of zero to all ranking results with $r_j^{best} > c$ (c.f. Section 2.2). Subsequently, it was shown that $h(\boldsymbol{r}_j)$ and its modifications do not only allow to quantify the data dredging potential of a single ranking result but also to rank the variables according to their data dredging potential and to quantify the overall data dredging potential. Ranking the variables with respect to their data dredging potential might be particularly relevant if the researcher wants to compare the ranking result of a specific variable X_{j*} (e.g. his/her "favourite" variable) with the ranking results of the other variables. This enables the researcher to assess the rank of X_{j*} with respect to its data dredging potential and to check if X_{j*} is among the variables with the highest data dredging potential. For this purpose, it was recommended to use a combination of $h(\boldsymbol{r}_j; \alpha)$ and a cutoff value c. Even if there is no specific variable that the researcher is interested in, $h(\boldsymbol{r}_j)$ can be used to quantify the overall data dredging potential of the ranking results. The idea behind this procedure is to provide an overview about how unstable the ranking results are with respect to data dredging. In order to quantify the overall data dredging potential, an aggregated version of $h(\boldsymbol{r}_j)$, $H(c)$, was proposed. It can be interpreted as the mean difference between the average and the best rank over all variables with $r_j^{best} \leqslant c$. The use of $H(c; \alpha)$, which is an aggregated version of $h(\boldsymbol{r}_j; \alpha)$, was not recommended to use in practice because of its less intuitive interpretation (c.f. Section 2.3).

The proposed framework was illustrated in the context of gene rankings using simulated (Chapter 3) and real data sets (Chapter 4 and 5). Regarding the outcome of interest, two scenarios were considered: Data with binary outcome (Chapter 3 and 4) and data with survival outcome (Chapter 5). Correspondingly, the data sets with binary outcome were ranked by other methods than the data set with survival outcome.

As proposed in Chapter 2, the variables in each data set were ranked with respect to their data dredging potential. For this purpose, the data dredging potential of each ranking

result was quantified by a combination of $\alpha \in \{0, 0.5, 0.8\}$ and $c = 100$. On average, the best ranks of the top-10 ranking results with the highest data dredging potential were higher for $\alpha = 0$ than for $\alpha > 0$. This result was expected since for $\alpha > 0$, smaller ranks are considered as more relevant for data dredging, whereas for $\alpha = 0$, the actual value of r_j^{best} is not taken into account.
The overall data dredging potential in each data set was assessed in three steps. First, the distribution of each ranking result and the number of variables with $r_j^{best} \leqslant c$ were considered. In the second step, the overall data dredging potential was assessed by $H(c)$, which quantifies the mean variability in ranking results with respect to r_j^{best}. In general, the results of step 2 were consistent with the results of step 1: In the simulation, the overall data dredging potential increased with decreasing number of observations and increasing number of variables. The correlation structure did not appear to affect the overall data dredging potential. Regarding the real data applications, the overall data dredging potential was considerably smaller if the number of methods was reduced. As a threshold for unproblematic data dredging potential, a value of $H(c) = 25$ was proposed. Taking this as a basis, almost all data sets show a problematic data dredging potential. As a third step, the methods were analysed with respect to their contribution to the overall data dredging potential. In the two-group setting (simulated and real data), the fold-change method was shown to yield the highest increase in data dredging potential when added to the set of ranking methods. For the survival outcome, the methods SIS and IPOD increased the overall data dredging potential considerably.

To summarize, the framework proposed in this thesis allows the researcher to get a concrete idea of how unstable ranking results are with respect to data dredging. The importance of acknowledging data dredging potential becomes drastically apparent regarding the high variability in ranking results with respect to r_j^{best} that was observed in the data applications. Quantifying the data dredging potential in ranking results could help to raise awareness of data dredging, which is a practice that can lead to a substantial optimistic bias. In contrast to existing stability measures for ranking lists, the provided framework allows to quantify the variability in ranking results with respect to r_j^{best} (i.e. the result that is relevant for data dredging). The proposed framework is flexible in the sense that by choosing the parameters α and c adequately, $h(\boldsymbol{r}_j)$ can be individually adjusted to the respective research question. Specifically, this means that the framework could also be applied to other research fields than the selection of biomarkers. Moreover, the framework can not only be used to compare the results of different ranking criteria (as shown in the applications) but also to compare ranking results that result of all possible choices concerning the analysis strategy. Another advantage of the framework is that it enables the researcher to compare the overall data dredging potential resulting from different data sets or different numbers of ranking methods. For example, this allows to study the impact of

the number of variables, observations and methods on the overall data dredging potential. As already stated in the thesis, the flexibility of the framework regarding the parameters α and c can also be seen as a drawback. Although some general recommendations were given, it might still be difficult to choose an appropriate value for α and c since there is no objective criterion for this choice. Another drawback of the framework is that it is difficult to define a generally valid threshold that allows to differentiate between problematic and unproblematic (overall) data dredging potential. Thus, the aims of future research could be to establish more detailed guidelines regarding the choice of parameter α and c and the definition of problematic dredging potential.
Moreover, future research could investigate the association between data dredging potential and data dredging bias. As stated in the thesis, this would require simulated data since the true rank of each variable is in general not known. Furthermore, defining the data dredging bias would require a sound theoretical basis. For instance, it might not be appropriate to simply define the data dredging bias as the difference or ratio of true rank and best rank. Moreover, it could be reasonable to define the data dredging bias depending on the application. For example, in the case of differential expression analysis, it could be useful to differentiate between differentially expressed and non-differentially expressed genes when quantifying the data dredging bias.

Overall, the framework introduced in this thesis can be regarded as a first step towards acknowledging the variability in ranking results with respect to data dredging. Although more research is needed on this area, the framework can hopefully contribute to raising awareness of data dredging in ranking analyses.

References

Boulesteix, A.-L., Hornung, R., and Sauerbrei, W. (2017). On fishing for significance and statistician's degree of freedom in the era of big molecular data. In Pietsch, W., Wernecke, J., and Ott, M., editors, *Berechenbarkeit der Welt?*, pages 155–170. Springer, Wiesbaden.

Boulesteix, A.-L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in bioinformatics*, 10(5):556–568.

Boulesteix, A.-L. and Strobl, C. (2009). Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC medical research methodology*, 9:85.

Chen, X., Chen, X., and Wang, H. (2018). Robust feature screening for ultra-high dimensional right censored data via distance correlation. *Computational Statistics & Data Analysis*, 119:118–138.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–2778.

Chu, G., Seo, M., Li, J., Narasimhan, B., Tibshirani, R., and Tusher, V. (2002). *SAM, Significance Analysis of Microarrays, Users guide and technical document.* R vignette, available at `http://statweb.stanford.edu/~tibs/SAM/sam.pdf`, last accessed on 11/06/19.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export Tables to LaTeX or HTML.* R package version 1.8-4.

Dessì, N., Pascariello, E., and Pes, B. (2013). A comparative analysis of biomarker selection techniques. *BioMed research international*, 2013:1–10.

Dudoit, S., Shaffer, J. P., and Block, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Edelmann, D., Hummel, M., Hielscher, T., and Benner, A. (2019a). *MVS: R Code used for Edelmann et al.* R package version 0.1.0, available at `https://github.com/thomashielscher/MVS`, last accessed on 10/25/19.

Edelmann, D., Hummel, M., Hielscher, T., Saadati, M., and Benner, A. (2019b). Marginal variable screening for survival endpoints. *Biometrical Journal*, pages 1–17.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. In Berger, J. O., Cai, T. T., and Johnstone, I. M., editors, *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, Institute of Mathematical Statistics Collections, pages 70–86. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2019). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-11.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Hong, H. G., Chen, X., Christiani, D. C., and Li, Y. (2018). Integrated powered density: Screening ultrahigh dimensional covariates with survival outcomes. *Biometrics*, 74(2):421–429.

Hu, Z.-Z., Huang, H., Wu, C. H., Jung, M., Dritschilo, A., Riegel, A. T., and Wellstein, A. (2011). Omics-based molecular target and biomarker identification. *Methods in molecular biology (Clifton, N.J.)*, 719:547–571.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Irigoien, I. and Arenas, C. (2018). Identification of differentially expressed genes by means of outlier detection. *BMC bioinformatics*, 19(1):317:1–317:20.

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., and Guedj, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PloS one*, 5(9):e12336.

Klau, S., Martin-Magniette, M.-L., Boulesteix, A.-L., and Hoffmann, S. (2019). Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. *Biometrical Journal*, pages 1–18.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398.

Li, X. (2018). *ALL: A data package*. R package version 1.24.0, available at `http://bioconductor.org/packages/release/data/experiment/html/ALL.html`, last accessed on 10/25/19.

Marín-Franch, I. (2018). Publication bias and the chase for statistical significance. *Journal of optometry*, 11(2):67–68.

Meschiari, S. (2015). *latex2exp: Use LaTeX Expressions in Plots*. R package version 0.4.0.

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2018). A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114(526):928–937.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rencher, A. C. (2002). *Methods of multivariate analysis*. Wiley series in probability and mathematical statistics. J. Wiley, New York, 2 edition.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W. C., and et al. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, 3(2):185–197.

Saldana, D. F. and Feng, Y. (2018). SIS: An R Package for Sure Independence Screening in Ultrahigh-Dimensional Statistical Models. *Journal of Statistical Software*, 83(2):1–25.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Slawski, M. and Boulesteix, A.-L. (2018). *GeneSelector: Stability and Aggregation of ranked gene lists*. R package version 2.30.0, available at `https://bioconductor.riken.jp/packages/3.8/bioc/html/GeneSelector.html`, last accessed on 11/05/19, note that package was removed with Bioconductor 3.9 release.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(3).

Szucs, D. (2016). A tutorial on hunting statistical significance by chasing n. *Frontiers in psychology*, 7:1444.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20. R package version 1.4.3.

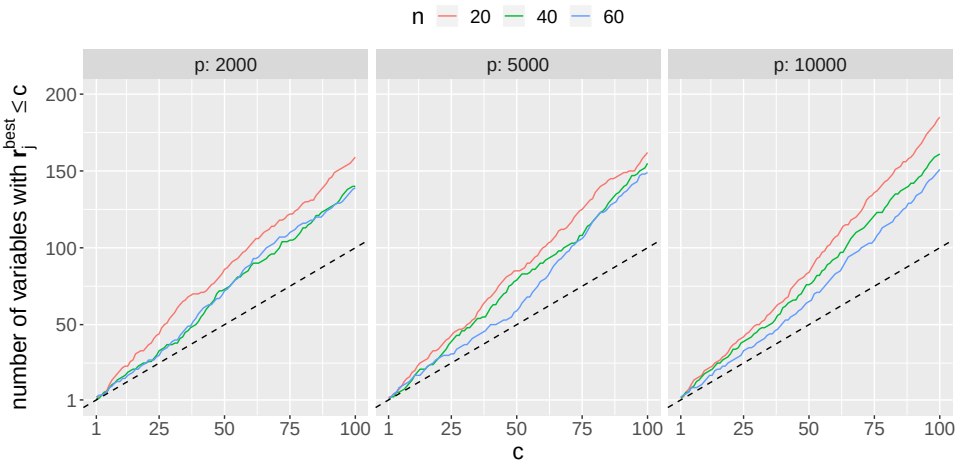
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. R package version 3.2.1.

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.

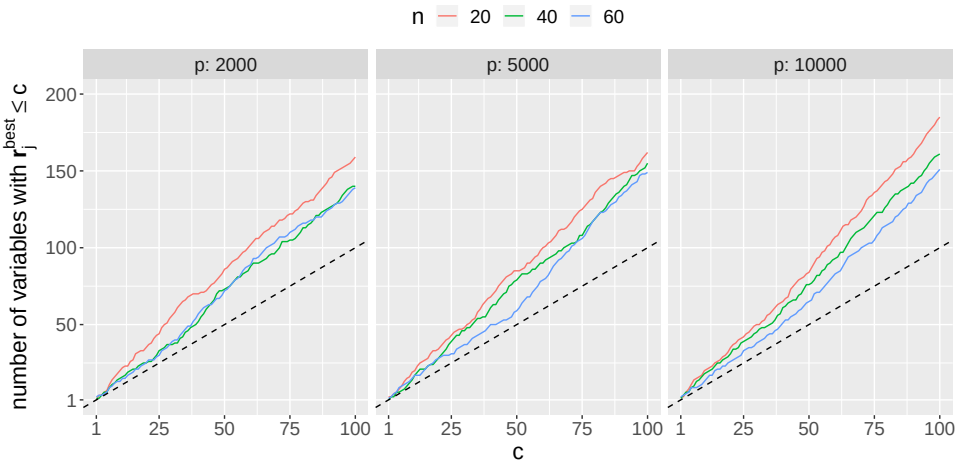
Wickham, H., François, R., Henry, L., and Müller, K. (2019). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3.

Wu, B. (2005). Differential gene expression detection using penalized linear regression models: the improved sam statistics. *Bioinformatics*, 21(8):1565–1571.

A. Additional figures and tables

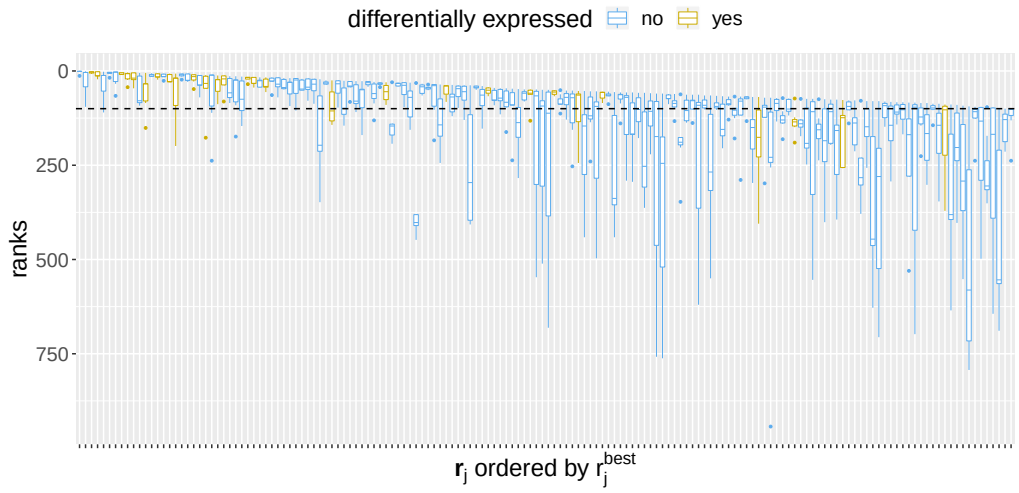


(a) $\rho = 0.4$

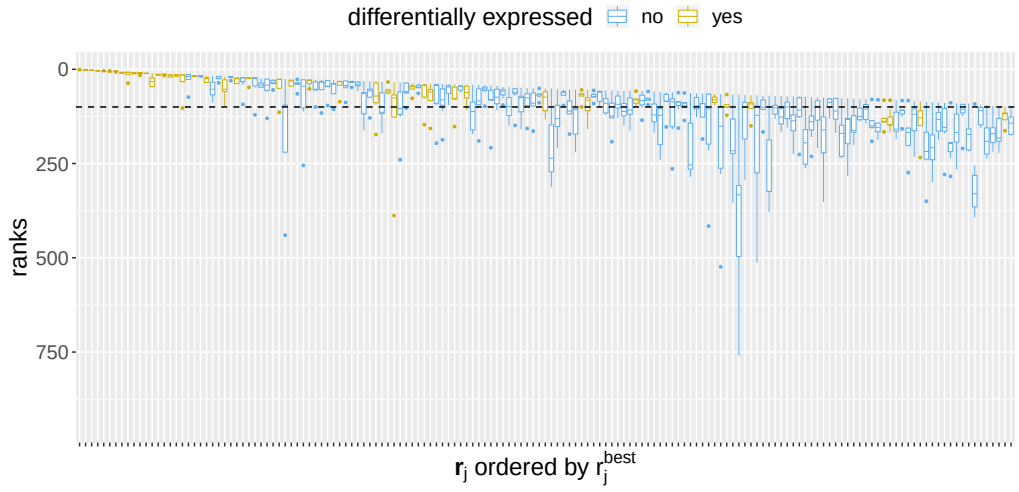


(b) $\rho = 0.8$

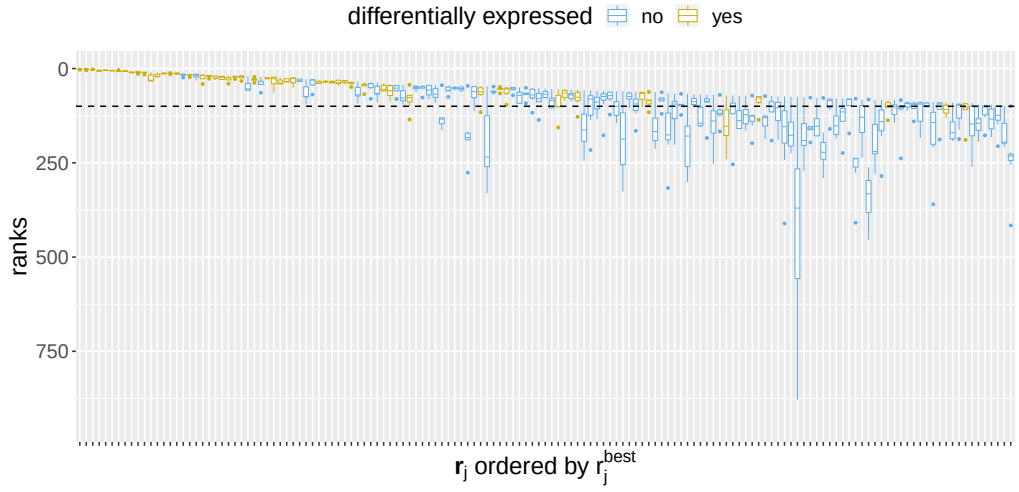
Figure A.1.: Number of variables with $r_j^{best} \leqslant c$, $c \in [1, 100]$, for simulated data sets with $\rho \in \{0.4, 0.8\}$. If all methods yield the same rank for each gene, the number of variables $r_j^{best} \leqslant c$ is equal to c (dotted line).



(a) $p = 5000, n = 20$

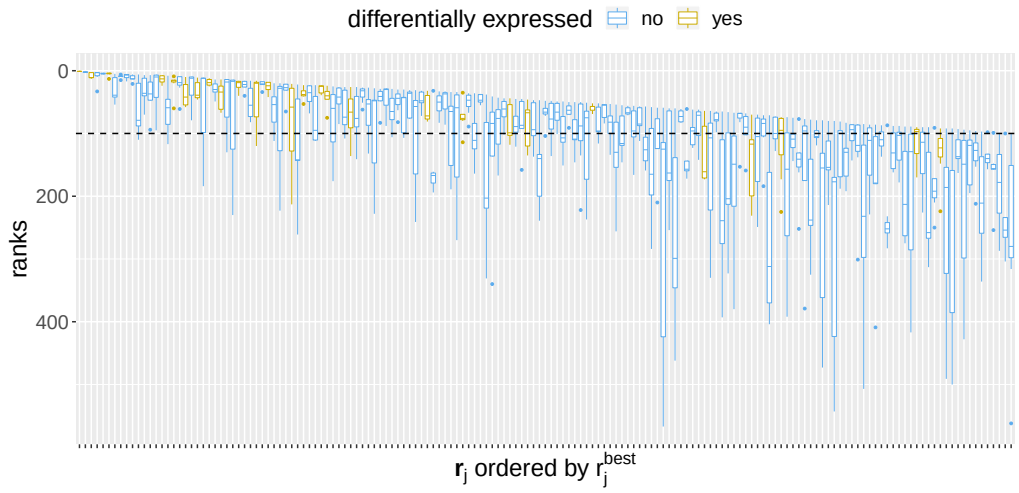


(b) $p = 5000, n = 40$

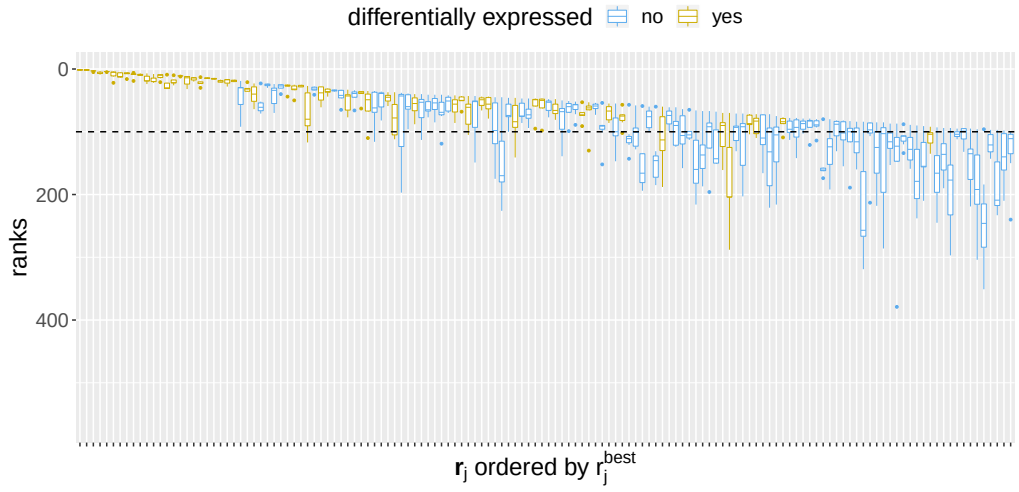


(c) $p = 5000, n = 60$

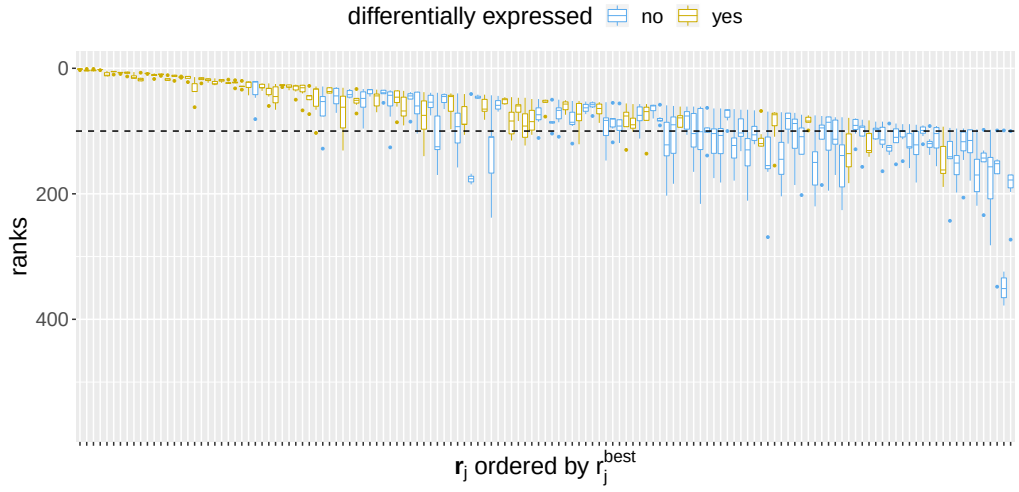
Figure A.2.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0$, $p = 5000$ and $n \in \{20, 40, 60\}$.



(a) $p = 2000, n = 20$

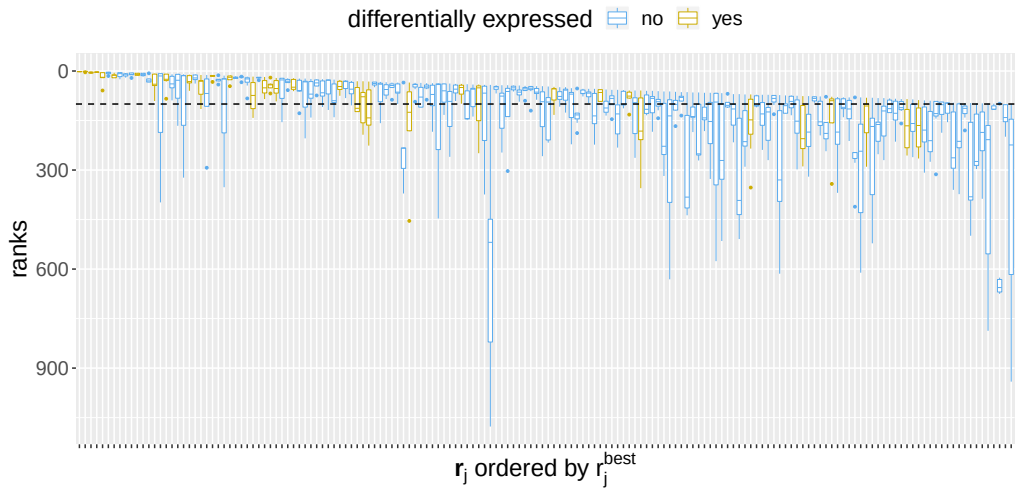


(b) $p = 2000, n = 40$

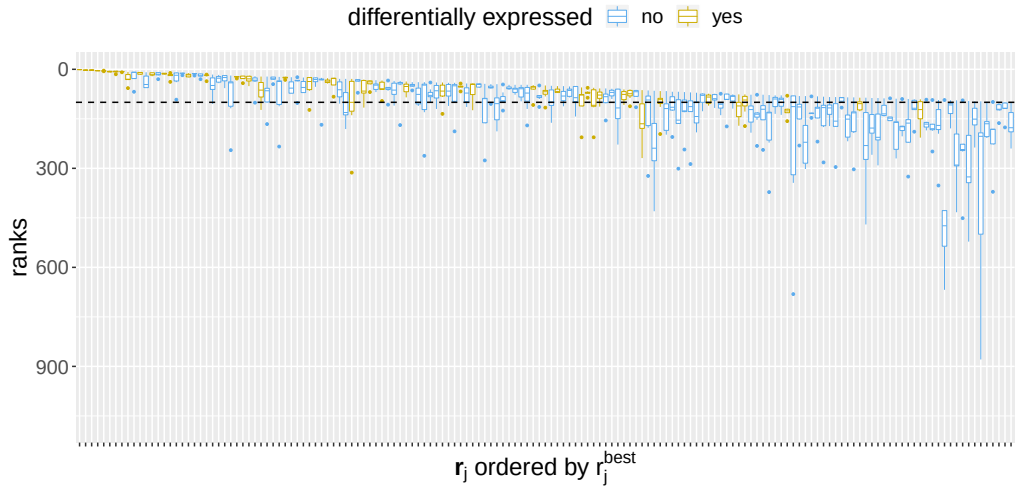


(c) $p = 2000, n = 60$

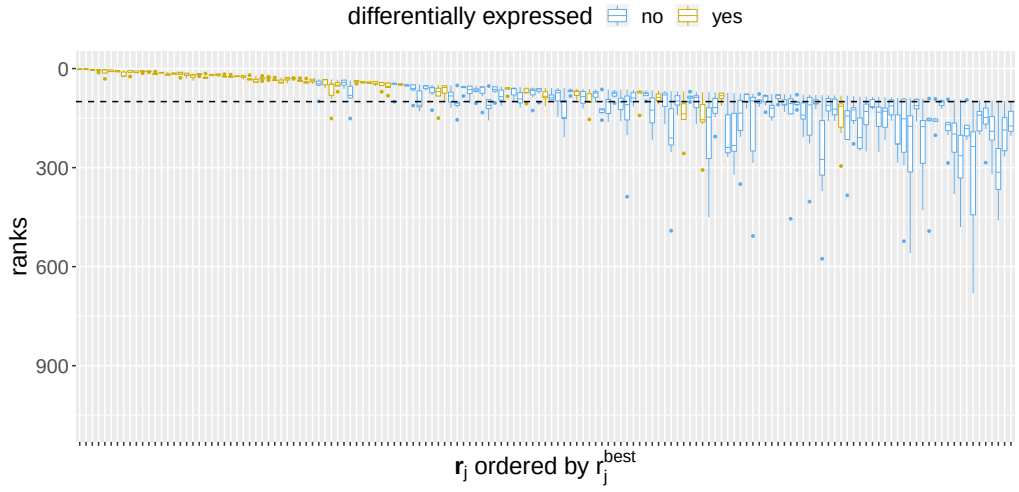
Figure A.3.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0.4$, $p = 2000$ and $n \in \{20, 40, 60\}$.



(a) $p = 5000, n = 20$

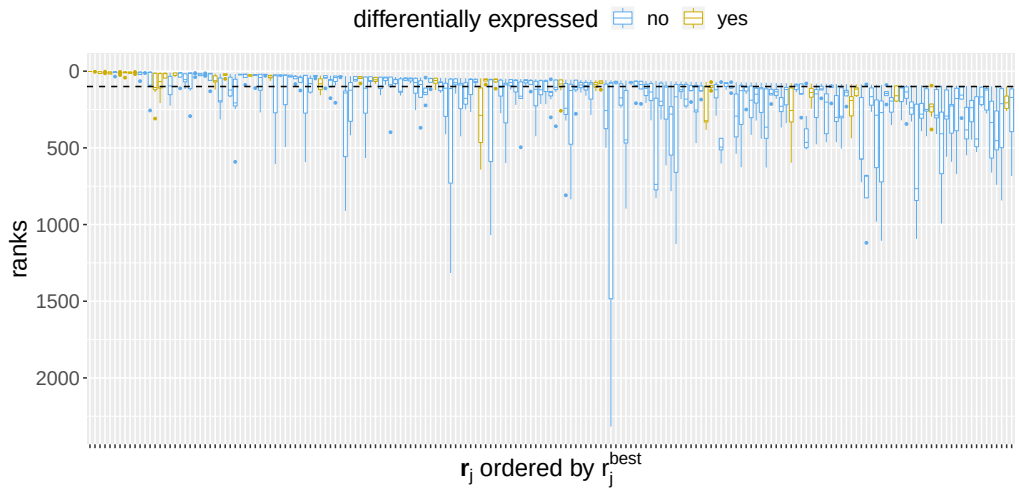


(b) $p = 5000, n = 40$

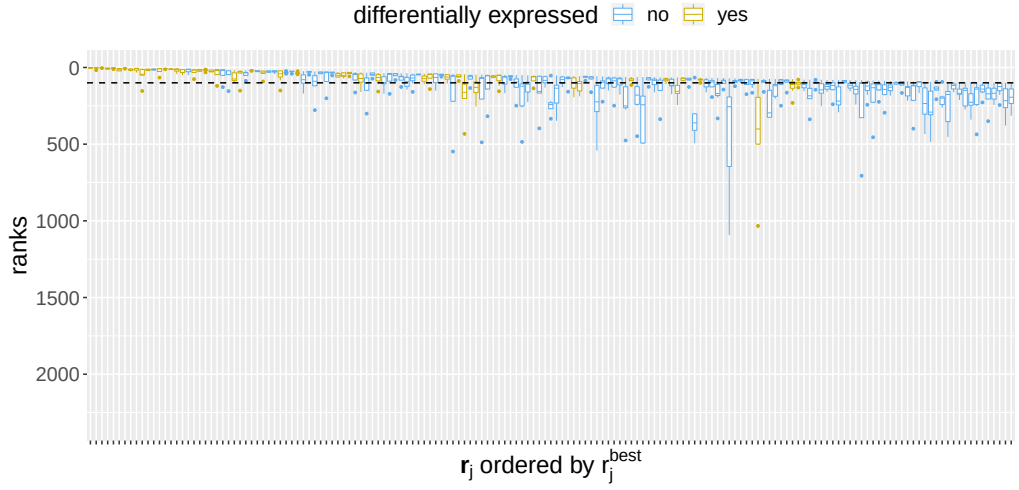


(c) $p = 5000, n = 60$

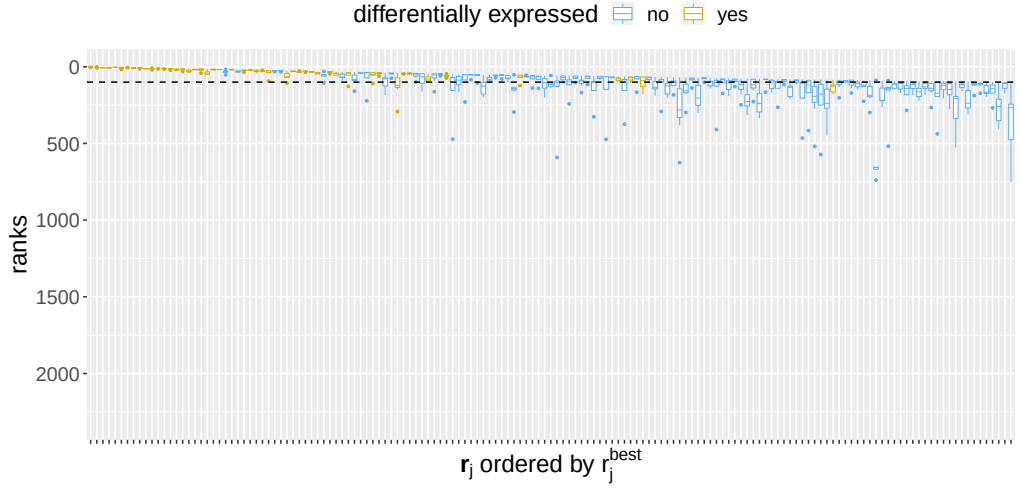
Figure A.4.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0.4$, $p = 5000$ and $n \in \{20, 40, 60\}$.



(a) $p = 10000, n = 20$

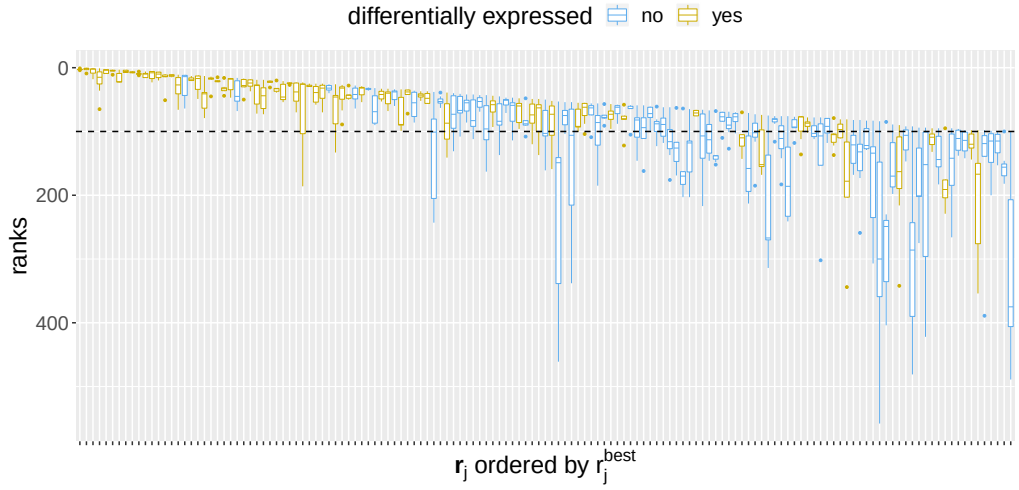


(b) $p = 10000, n = 40$

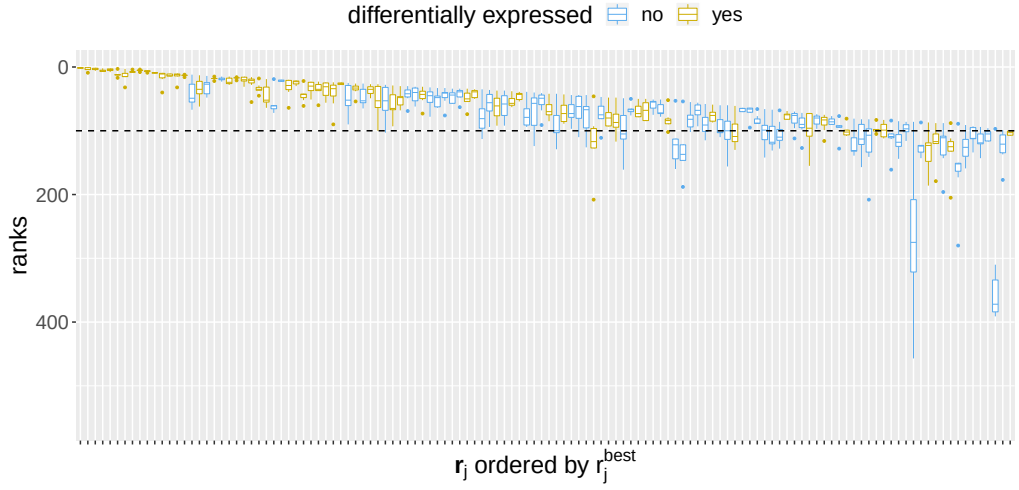


(c) $p = 10000, n = 60$

Figure A.5.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0.4$, $p = 10000$ and $n \in \{20, 40, 60\}$.



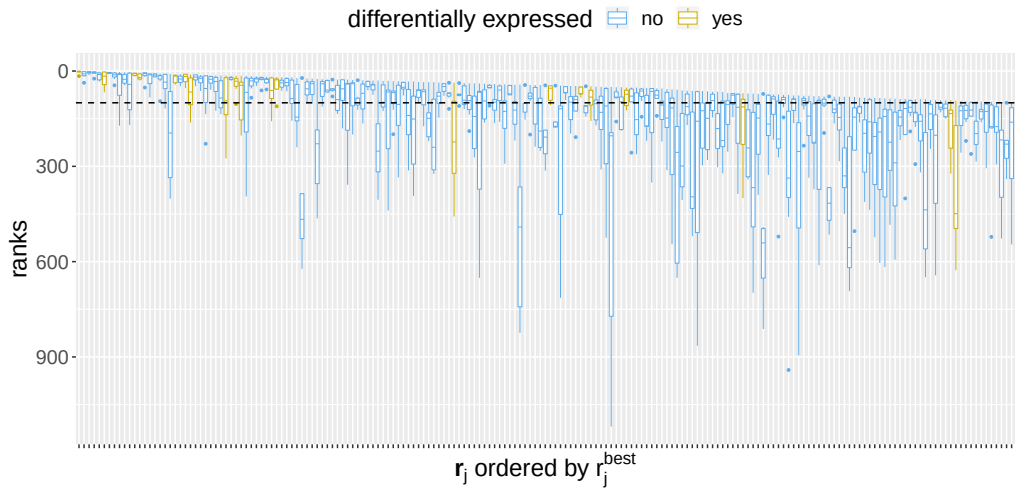
(a) $p = 2000, n = 20$



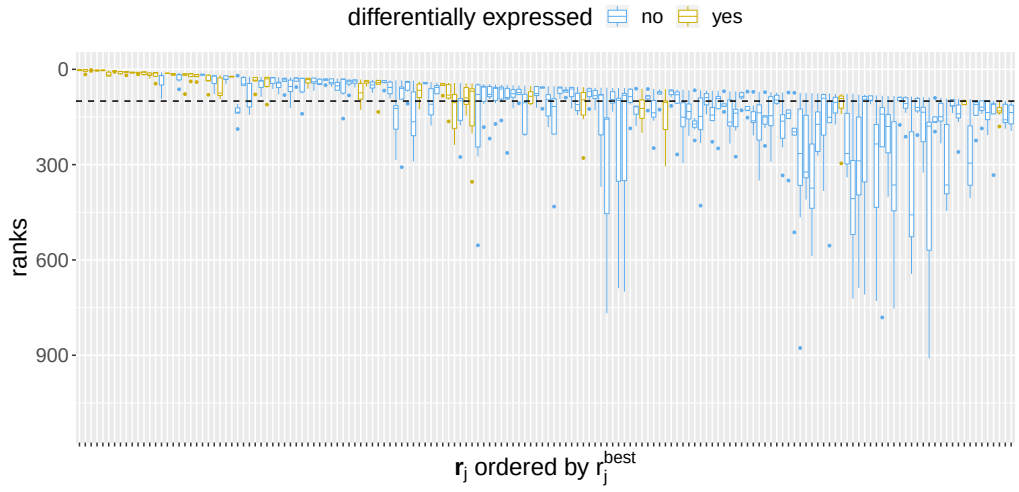
(b) $p = 2000, n = 40$

(c) $p = 2000, n = 60$

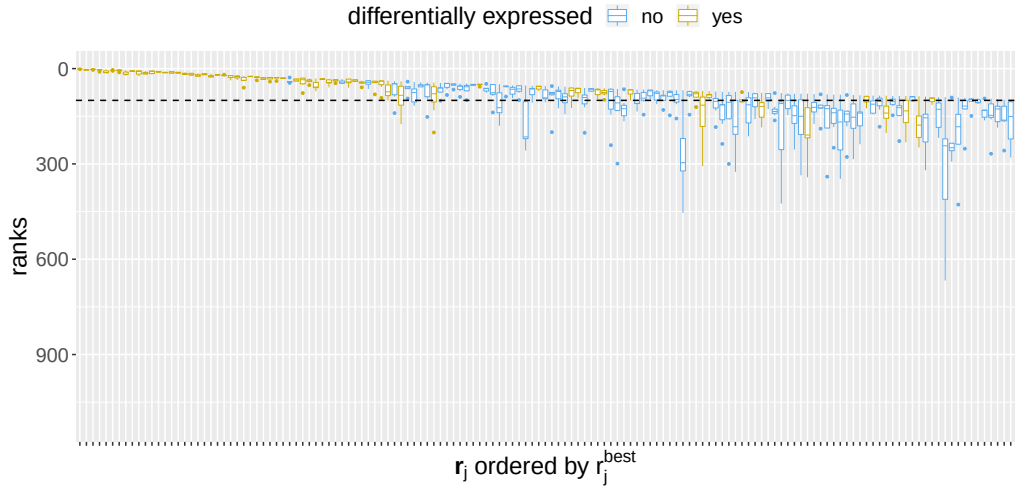
Figure A.6.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0.8$, $p = 2000$ and $n \in \{20, 40, 60\}$.



(a) $p = 5000, n = 20$

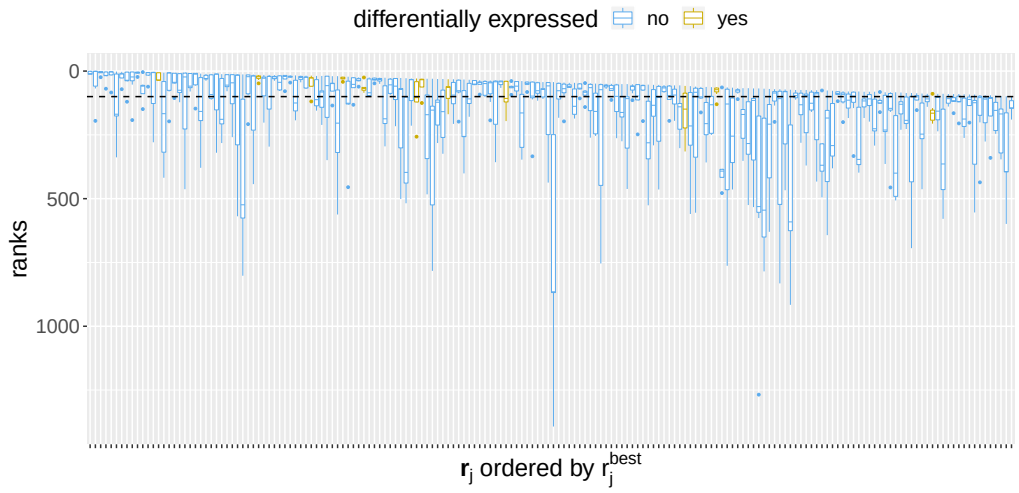


(b) $p = 5000, n = 40$

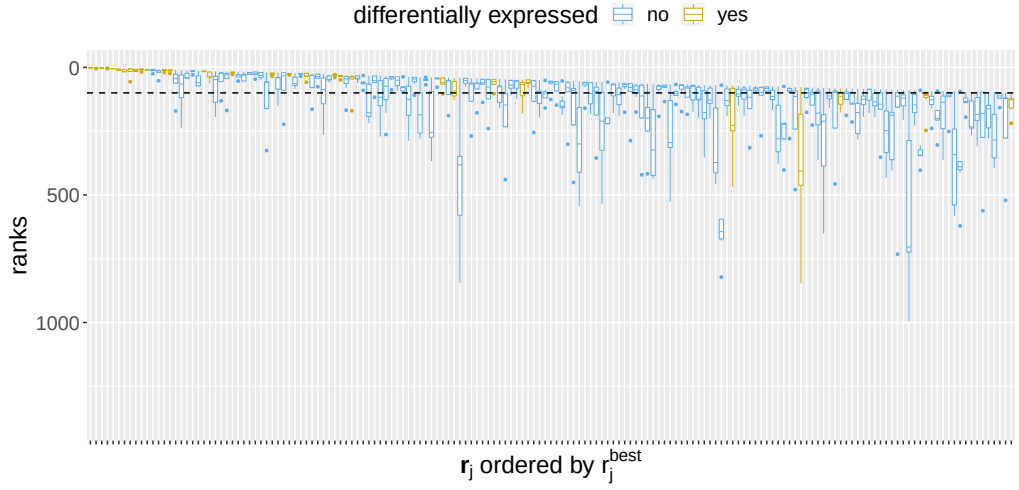


(c) $p = 5000, n = 60$

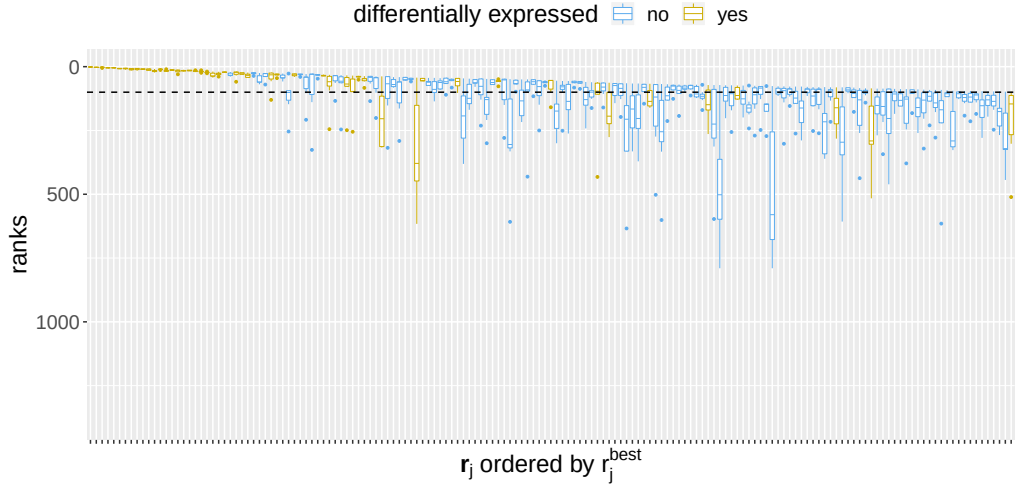
Figure A.7.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0.8$, $p = 5000$ and $n \in \{20, 40, 60\}$.



(a) $p = 10000, n = 20$



(b) $p = 10000, n = 40$



(c) $p = 10000, n = 60$

Figure A.8.: Boxplots showing the distribution of each $\boldsymbol{r}_j$ with $r_j^{best} \leqslant 100$ (dashed line) for three simulated datasets with $\rho = 0.8$, $p = 10000$ and $n \in \{20, 40, 60\}$.

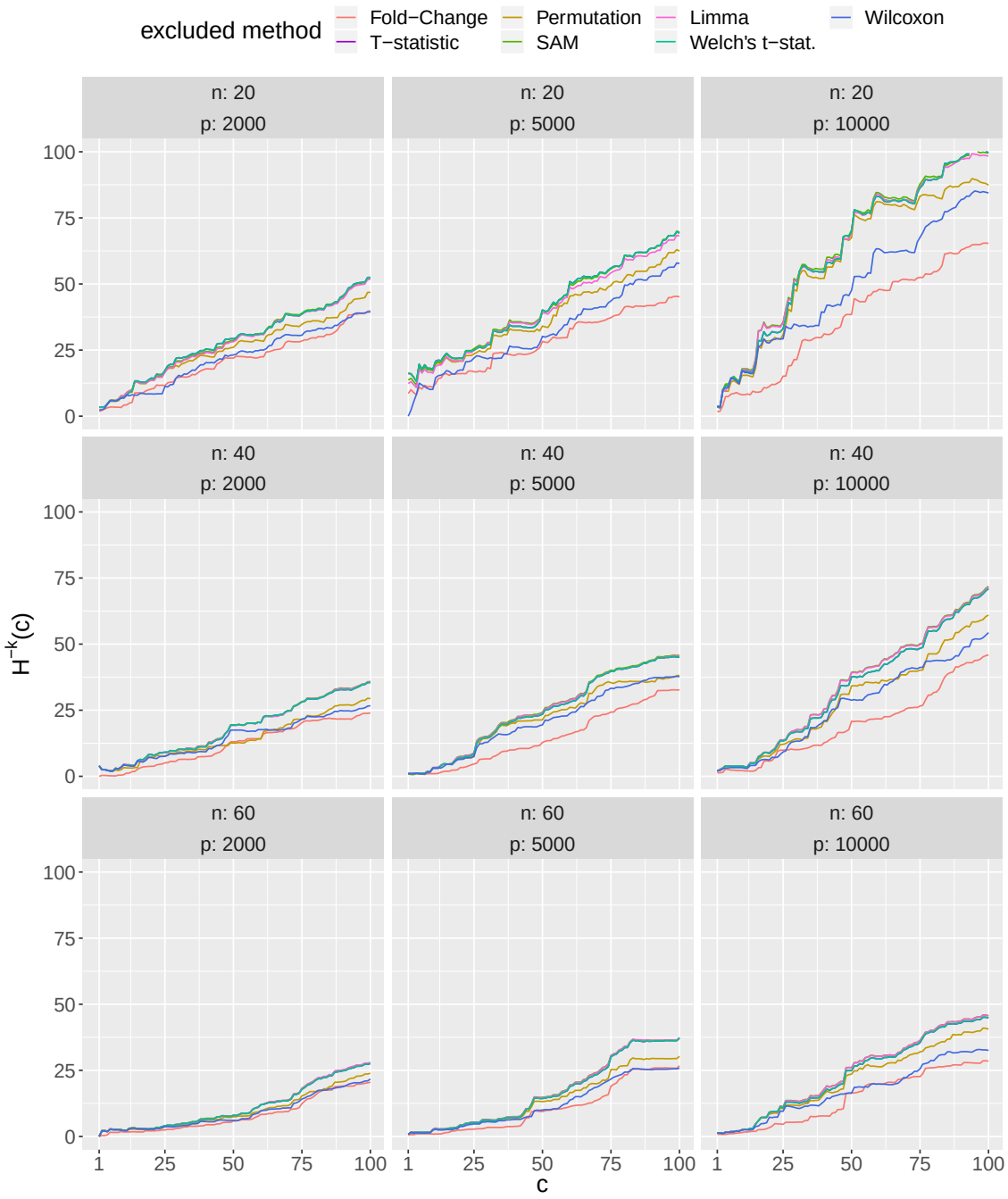


Figure A.9.: Overall data dredging potential that arises when the variables rankings are performed without usig method $k = 1, \ldots, 7$ is not considered. The method with the smallest value of $H^{-k}(c)$ is the method that yields the highest increase in overall data dredging potential when it is added to the set of the other ranking methods. The figure shows nine simulated data sets with $\rho = 0$, $p = \{2000, 5000, 10000\}$ and $n \in \{20, 40, 60\}$.

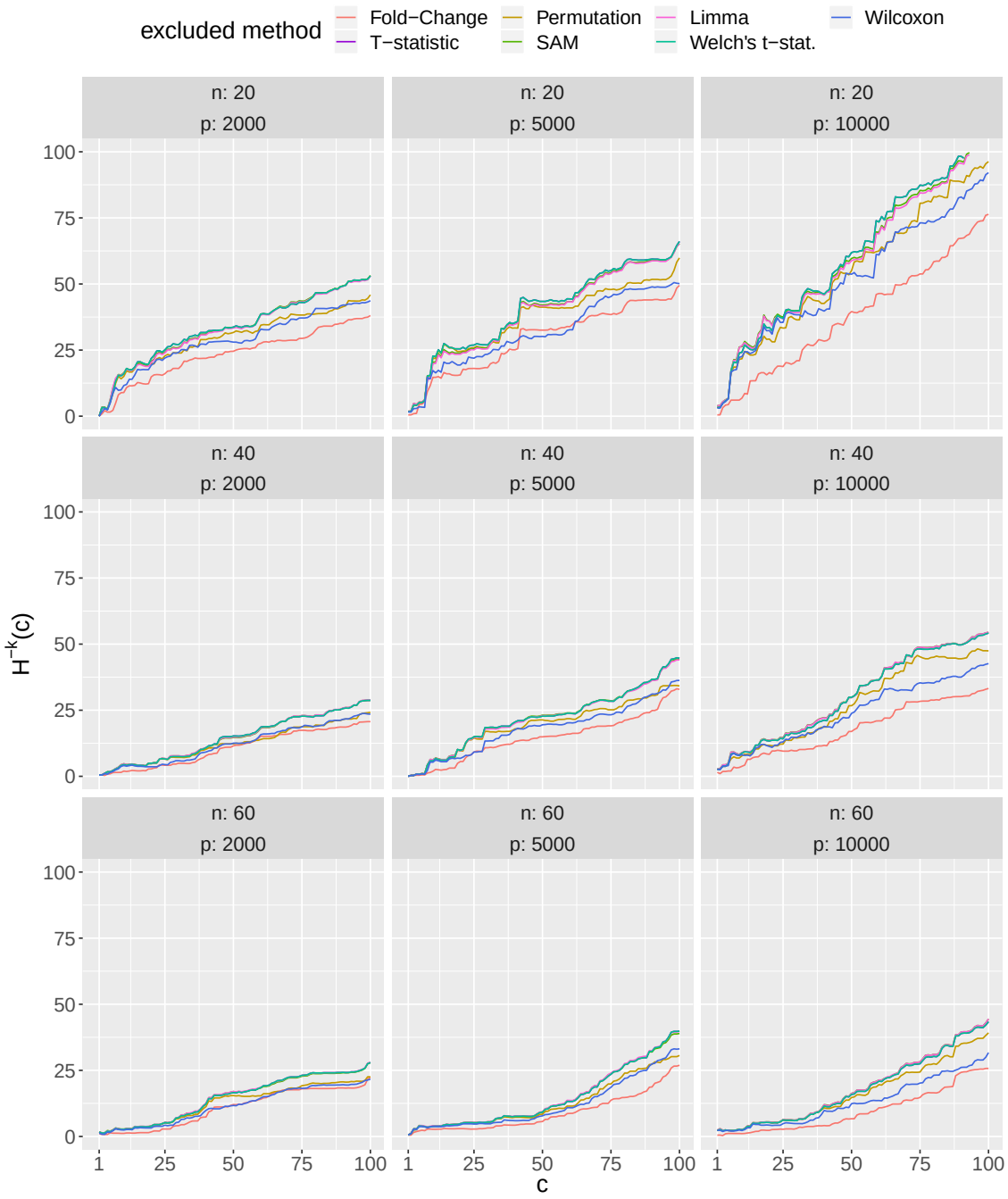


Figure A.10.: Overall data dredging potential that arises when the variables rankings are performed without usig method $k = 1, \ldots, 7$ is not considered. The method with the smallest value of $H^{-k}(c)$ is the method that yields the highest increase in overall data dredging potential when it is added to the set of the other ranking methods. The figure shows three simulated data sets with $\rho = 0.4$, $p \in \{2000, 5000, 10000\}$ and $n \in \{20, 40, 60\}$.

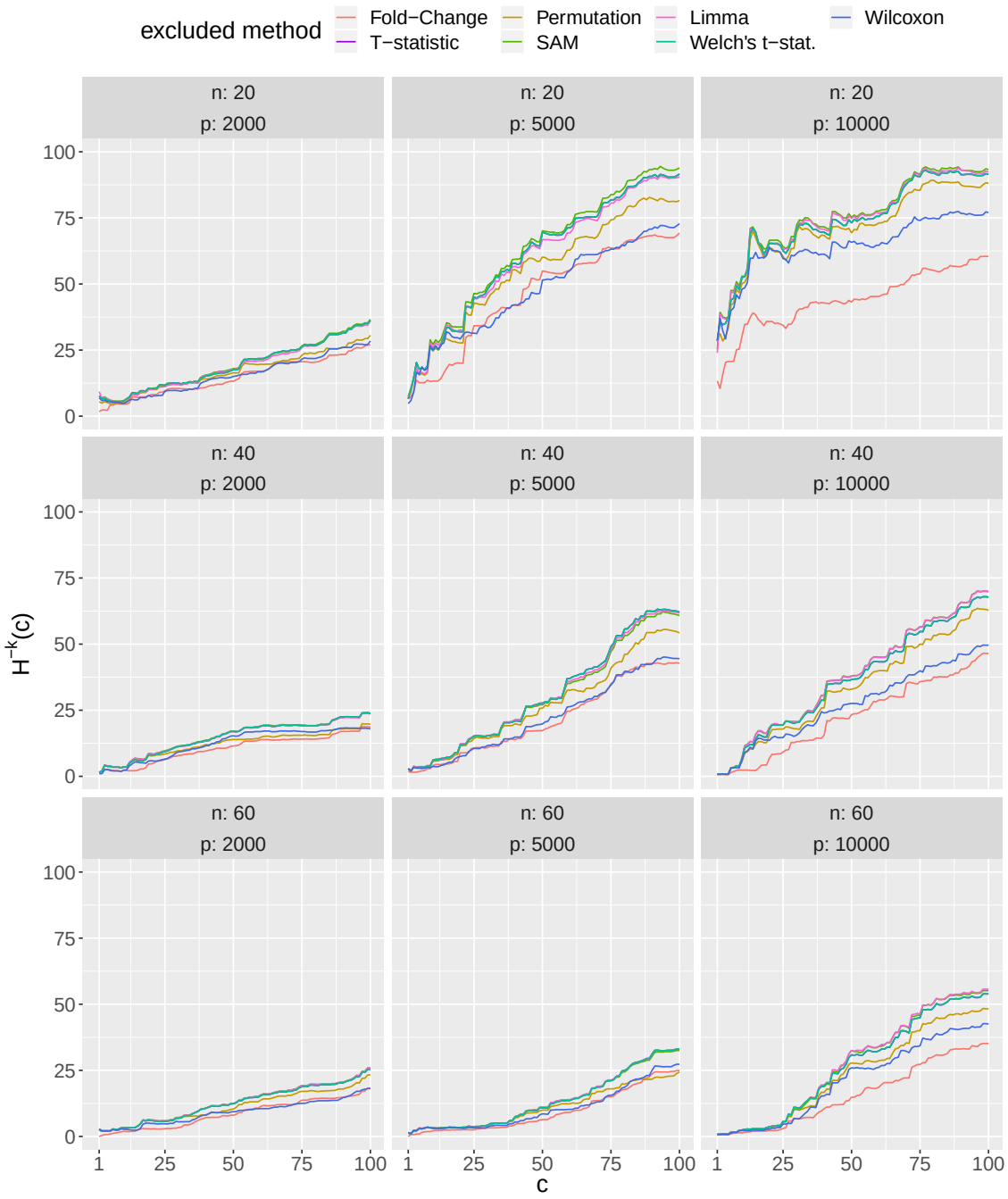


Figure A.11.: Overall data dredging potential that arises when the variables rankings are performed without usig method $k = 1, \ldots, 7$ is not considered. The method with the smallest value of $H^{-k}(c)$ is the method that yields the highest increase in overall data dredging potential when it is added to the set of the other ranking methods. The figure shows nine simulated data sets with $\rho = 0.8$, $p \in \{2000, 5000, 10000\}$ and $n \in \{20, 40, 60\}$.

Table A.1.: Top-10 lists of ranking results with the highest data dredging potential resulting from the ALL data set with $c = 100$, $\alpha \in \{0, 0.5, 0.8\}$ and $m = 4$. Empty cells denote that the ranking result is not among the top-10. For each ranking result, r_j^{best} is highlighted in red.

gene	$\boldsymbol{r}_j$	$h(\boldsymbol{r}_j;\alpha)$			rank (w.r.t. $h(\boldsymbol{r}_j;\alpha)$)		
		$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$
35192_at	(2244 1984 2214 100)	1535.50	153.55	38.57	1	3	
38004_at	(2020 1630 1982 90)	1340.50	141.30	36.63	2	5	
1914_at	(1990 1175 1932 61)	1228.50	157.29	45.83	3	1	8
1299_at	(1706 1313 1669 94)	1101.50	113.61	29.07	4		
2004_at	(1589 1065 1541 66)	999.25	123.00	35.00	5	8	
1974_s_at	(1426 1167 1401 48)	962.50	138.92	43.49	6	6	
1565_s_at	(1459 965 1421 83)	899.00	98.68	26.21	7		
712_s_at	(1098 1311 1109 76)	822.50	94.35	25.73	8		
40434_at	(888 1147 891 98)	658.00	66.47	16.80	9		
626_s_at	(930 741 910 34)	619.75	106.29	36.90	10		
1475_s_at	(699 493 681 9)	461.50	153.83	79.58		2	5
34098_f_at	(422 220 409 3)	260.50	150.40	108.17		4	2
1472_g_at	(484 253 461 5)	295.75	132.26	81.61		7	4
38124_at	(194 120 180 1)	122.75	122.75	122.75		9	1
38279_at	(627 471 611 13)	417.50	115.79	53.64		10	7
37185_at	(6 2 5 564)	142.25	100.59	81.70			3
1473_s_at	(489 242 465 8)	293.00	103.59	55.51			6
577_at	(213 144 206 4)	137.75	68.88	45.44			9
953_g_at	(125 76 116 2)	77.75	54.98	44.66			10

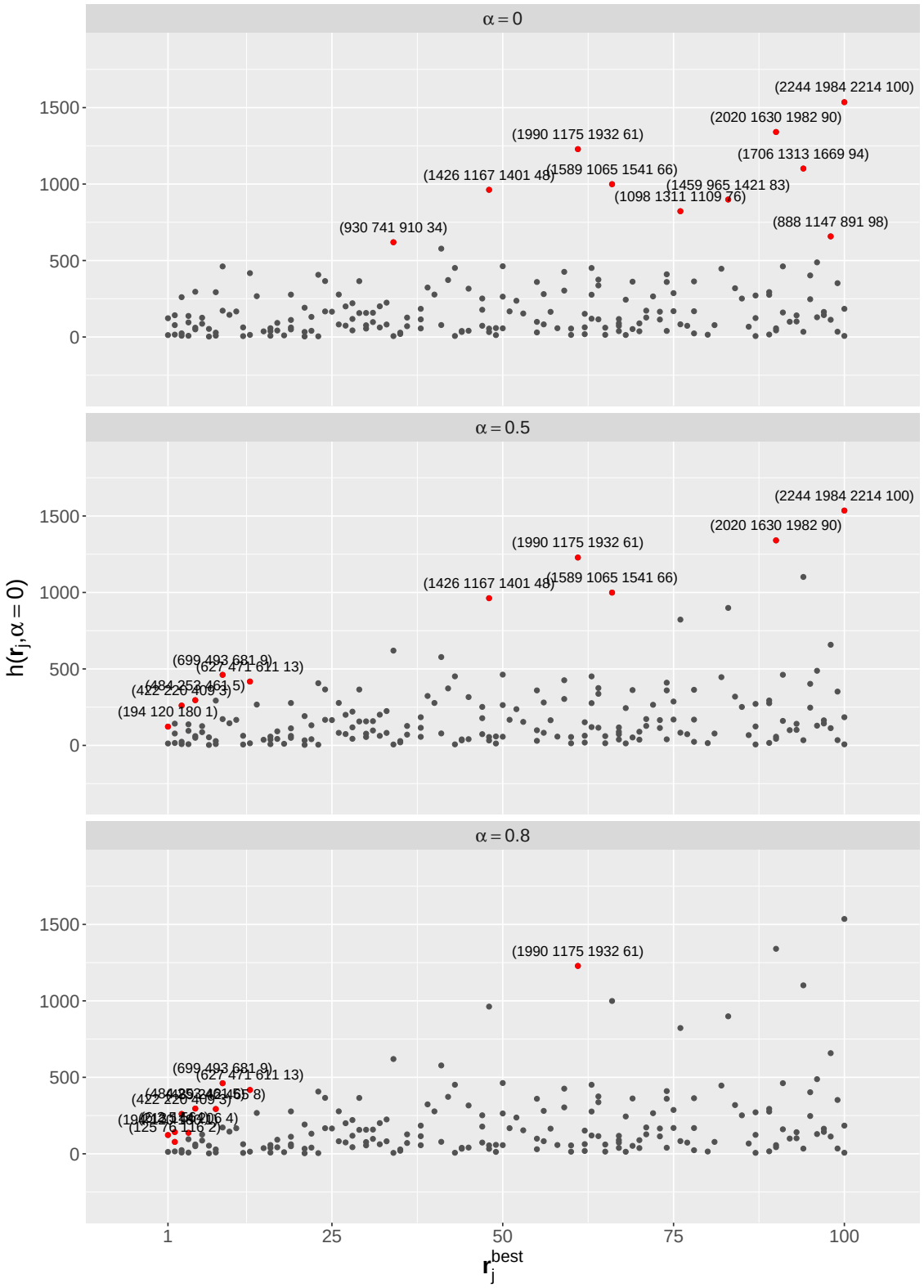


Figure A.12.: Data dredging potential of ranking results with $r_j^{best} \leqslant 100$ for $m = 4$. In each panel, the top-10 ranking results with the highest data dredging potential for $\alpha \in \{0, 0.5, 0.8\}$ are highlighted in red.

Table A.2.: Top-10 lists of ranking results with the highest data dredging potential resulting from the MCL data set with $c = 100$, $\alpha \in \{0, 0.5, 0.8\}$ and $m = 4$. Empty cells denote that the ranking result is not among the top-10. For each ranking result, r_j^{best} is highlighted in red.

gene	$\boldsymbol{r}_j$	$h(\boldsymbol{r}_j;\alpha)$			rank (w.r.t. $h(\boldsymbol{r}_j;\alpha)$)		
		$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$
16311	(918 57 51 77)	224.75	31.47	9.67	1	1	1
17545	(611 84 85 143)	146.75	16.01	4.24	2	2	5
24735	(512 117 89 75)	123.25	14.23	3.90	3	3	6
24758	(467 92 76 141)	118.00	13.54	3.69	4	4	7
24897	(325 90 111 229)	98.75	10.41	2.70	5	8	
25949	(77 200 242 156)	91.75	10.46	2.84	6	7	
27682	(94 214 217 214)	90.75	9.36	2.40	7		
33531	(69 256 170 133)	88.00	10.59	2.97	8	6	
27395	(73 224 198 104)	76.75	8.98	2.48	9		
17795	(45 152 178 111)	76.50	11.40	3.64	10	5	8
17069	(41 129 154 96)	64.00	10.00	3.28		9	10
28454	(54 131 193 118)	70.00	9.53	2.88		10	
28990	(24 1 1 3)	6.25	6.25	6.25			2
32187	(34 3 4 20)	12.25	7.07	5.09			3
17123	(70 22 13 7)	21.00	7.94	4.43			4
30282	(6 23 26 26)	14.25	5.82	3.40			9

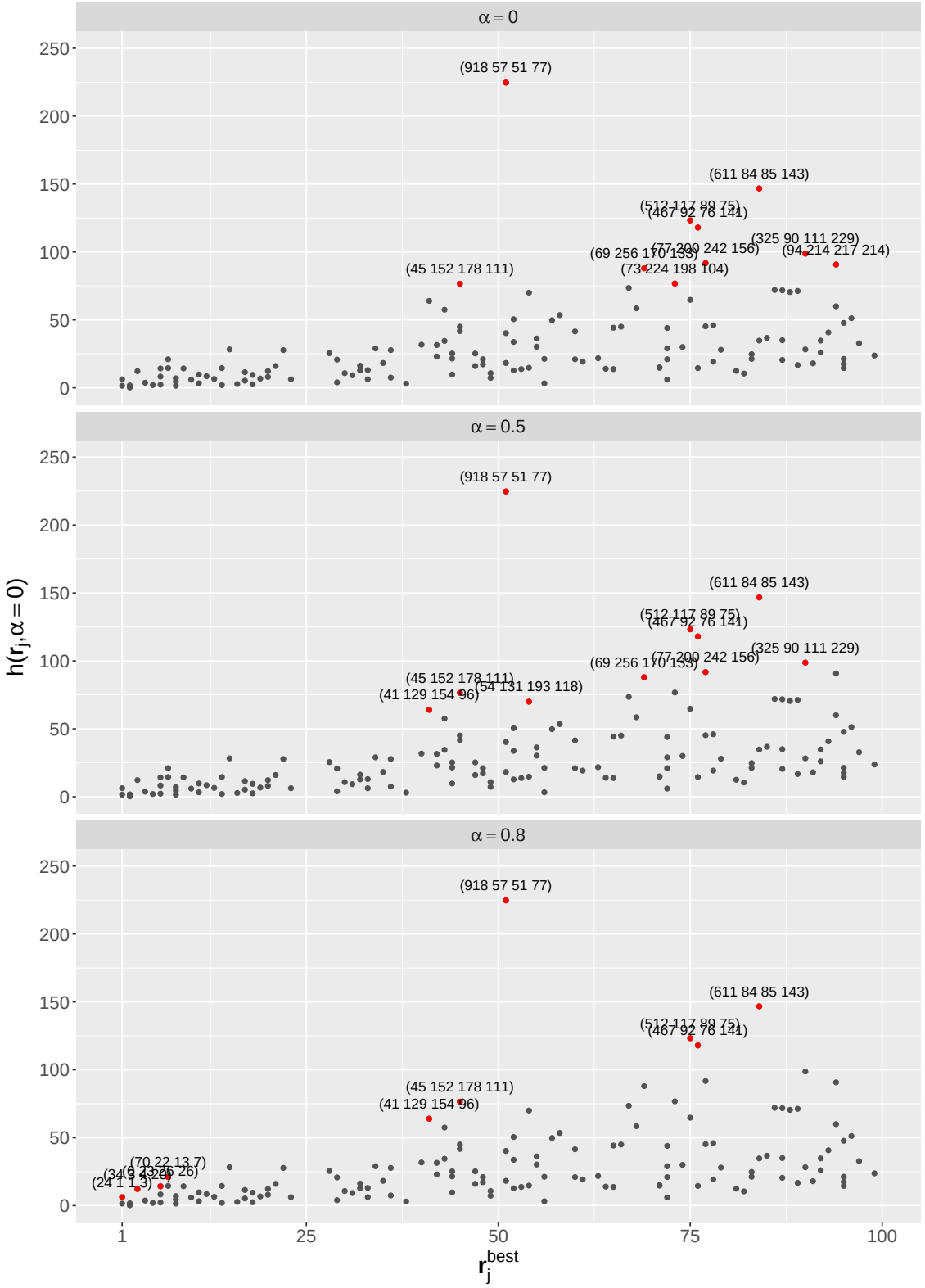


Figure A.13.: Data dredging potential of ranking results with $r_j^{best} \leqslant 100$ for $m = 4$. In each panel, the top-10 ranking results with the highest data dredging potential for $\alpha \in \{0, 0.5, 0.8\}$ are highlighted in red.

B. Electronic appendix

The electronic appendix comprises an electronic version of this thesis (`MA_Niessl.pdf`) as well as two folders. The first folder contains the additional figures and tables of the simulation that are not included in Appendix A (`Additional figures and tables`). The second folder (`R Code`) consists of four subfolders that contain the code that produces the results shown in each chapter (`01 Framework`, `02 Simulation`, `03 Application Binary Outcome`, `04 Application Survival Outcome`).

Declaration of Authorship

I hereby confirm that I have authored this Master's thesis independently and without use of other resources than those indicated. The ideas taken directly or indirectly from external sources are duly acknowledged in the text. The material, either in full or in part, has not been previously submitted for grading at this or any other academic institution.

Munich, November 27, 2019

.....
Christina Nießl